FML_ACCIDENTS _ASSIGNMENT

2023-10-16

#Summary: 1. Our goal here is to predict whether an accident just reported will involve an injury (MAX_SEV_IR = 1 or 2) or will not (MAX_SEV_IR = 0). For this purpose, created a dummy variable called INJURY that takes the value "yes" if MAX_SEV_IR = 1 or 2, and otherwise "no. #Reasons: A dataset of automobile accidents is analyzed to predict whether a newly reported accident will result in an injury (INJURY = Yes) or not (INJURY = No). The code accomplishes the following: The function creates a binary dummy variable 'INJURY' which has the value"Yes" if 'MAX_SEV_IR' is either 1 or 2, otherwise it has the value "No". It calculates the proportion of accidents in the dataset that resulted in an injury (INJURY = Yes). This proportion is used as a threshold for making predictions. Based on the calculated percentage, it predicts whether there will be an injury for a newly reported accident with no further information. A higher proportion of injuries indicates a higher likelihood of injury. If the proportion of injuries is greater than 50%, the prediction is "Yes." Otherwise, the prediction is "No," suggesting a lower likelihood of injury. 2. Probability of injury was found to be 50.88%. 3. Create a pivot table that examines INJURY as a function of the two predictors WEATHER_R and TRAF_CON_R for the first 24 records. 4. Classified the 24 accidents using these probabilities and a cutoff of 0.5. 5. Bayes Probability found to be: [1] 0.6666667 0.1818182 0.0000000 0.0000000 0.0000000 1.0000000 [1] 0.3333333 0.8181818 1.0000000 1.0000000 1.0000000 0.0000000

6. Manual Naive Bayes Conditional Probability (Injury = Yes | Weather_R = 1, TRAF_CON_R = 1): 0
7. #RUNNING A NAIVE BAYES CLASSIFIER ON THE 24 RECORDS AND TWO PREDICTORS. #NOW,WE HAVE TO CHECK THE MODEL OUTPUT TO OBTAIN PROBABILITIES AND CLASSIFCATIONS FOR ALL 24 RECORDS. #AND THEN, WE ARE COMPARING TO BAYES CLASSIFCATION TO SEE IF THE RESULTING CLASSIFICATIONS ARE EQUIVALENT OR NOT.
8. Let us now return to the entire dataset. Partition the data into training (60%) and validation (40%)
9. overall error of the validation set Found to be: 0.477596

#Problem Statement

The file accidentsFull.csv contains information on 42,183 actual automobile accidents in 2001 in the United States that involved one of three levels of injury: NO INJURY, INJURY, or FATALITY. For each accident, additional information is recorded, such as day of week, weather conditions, and road type. A firm might be interested in developing a system for quickly classifying the severity of an accident based on initial reports and associated data in the system (some of which rely on GPS-assisted reporting).

Our goal here is to predict whether an accident just reported will involve an injury (MAX_SEV_IR = 1 or 2) or will not (MAX_SEV_IR = 0). For this purpose, create a dummy variable called INJURY that takes the value "yes" if MAX_SEV_IR = 1 or 2, and otherwise "no."

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?
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. Create a pivot table that examines INJURY as a function of the two predictors for these 12 records. Use all three variables in the pivot table as rows/columns.

- Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors.
- Classify the 24 accidents using these probabilities and a cutoff of 0.5.
- Compute manually the naive Bayes conditional probability of an injury given WEATHER_R = 1 and TRAF_CON_R = 1.
- 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?

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

- Run a naive Bayes classifier on the complete training set with the relevant predictors (and INJURY as the response). Note that all predictors are categorical. Show the confusion matrix.
- What is the overall error of the validation set?

#library

```
library(e1071)
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

#Import Data

```
Accidents_Data <- read.csv("C://Users//Princy//Documents//accidentsFull.csv")
head(Accidents_Data)
```

```
##   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
##   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
##   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
##   FATALITIES MAX_SEV_IR
## 1          0          1
## 2          0          0
## 3          0          0
## 4          0          0
## 5          0          0
## 6          0          1
```

#Create and insert a dummy variable called “INJURY” in the data.

```r
Accidents_Data$INJURY <- ifelse(Accidents_Data$MAX_SEV_IR>0, "yes", "no")

for (i in 1:dim(Accidents_Data)[2]) {
 if (is.character(Accidents_Data[, i])) {
 Accidents_Data[, i] <- as.factor(Accidents_Data[, i])
 }
}
head(Accidents_Data, 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
```

QUESTION-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?

#CREATING A TABLE BASED ON INJURY.

```
Injury_Table <- table(Accidents_Data$INJURY)
show(Injury_Table)
```

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

"{r} #CALUCATING THE PROBABILITY OF THE INJURY

```
Injury_Probablilty =
scales::percent(Injury_Table["yes"]/(Injury_Table["yes"]+Injury_Table["no"]),
0.01)
Injury_Probablilty
```

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

QUESTION-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.

```
#create a new subset with only the required records

Accidents_Data24 <- Accidents_Data[1:24, c('INJURY','WEATHER_R','TRAF_CON_R')]
Accidents_Data24
```

```
##    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
## 7      no         2          0
## 8     yes         1          0
## 9      no         2          0
## 10     no         2          0
## 11     no         2          0
## 12     no         1          2
## 13    yes         1          0
## 14     no         1          0
## 15    yes         1          0
## 16    yes         1          0
## 17     no         2          0
## 18     no         2          0
## 19     no         2          0
## 20     no         2          0
## 21    yes         1          0
## 22     no         1          0
## 23    yes         2          2
## 24    yes         2          0
```

#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(Accidents_Data24)
dt2 <- ftable(Accidents_Data24 [,-1])

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
```

Question-2(1)

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

```
#QUESTION4
#COMPUTING THE BAYES CONDITIONAL PROBABLITIES OF AN INJURY (INJURY = Yes) GIVEN THE SIX POSSIBILE COMBI

# Injury = yes

Prob1 = dt1[3,1] / dt2[1,1] # Injury, Weather=1 and Traf=0
Prob2 = dt1[4,1] / dt2[2,1] # Injury, Weather=2, Traf=0
Prob3 = dt1[3,2] / dt2[1,2] # Injury, W=1, T=1
Prob4 = dt1[4,2] / dt2[2,2] # I, W=2,T=1
Prob5 = dt1[3,3] / dt2[1,3] # I, W=1,T=2
Prob6 = dt1[4,3]/ dt2[2,3] #I,W=2,T=2
print(c(Prob1,Prob2,Prob3,Prob4,Prob5,Prob6))
```

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

```
# 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(N1,N2,N3,N4,N5,N6))
```

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

QUESTION-2(2)

#CLASSIFYING THE 24 ACCIDENTS USING THESES PROBABLITIES AND CUTOFF OF 0.5 #ADDING PROBABILITY RESULTS TO THE SUBSET

```
prob.inj <- rep(0,24)
for (i in 1:24) {
 print(c(Accidents_Data24$WEATHER_R[i],Accidents_Data24$TRAF_CON_R[i]))
 if (Accidents_Data24$WEATHER_R[i] == "1") {
 if (Accidents_Data24$TRAF_CON_R[i]=="0"){
 prob.inj[i] = Prob1
 }
 else if (Accidents_Data24$TRAF_CON_R[i]=="1") {
 prob.inj[i] = Prob3
 }
 else if (Accidents_Data24$TRAF_CON_R[i]=="2") {
 prob.inj[i] = Prob5
 }
 }
 else {
 if (Accidents_Data24$TRAF_CON_R[i]=="0"){
```

```
 prob.inj[i] = Prob2
 }
 else if (Accidents_Data24$TRAF_CON_R[i]=="1") {
 prob.inj[i] = Prob4
 }
 else if (Accidents_Data24$TRAF_CON_R[i]=="2") {
 prob.inj[i] = Prob6
 }
 }
}
```

```
## [1] 1 0
## [1] 2 0
## [1] 2 1
## [1] 1 1
## [1] 1 0
## [1] 2 0
## [1] 2 0
## [1] 1 0
## [1] 2 0
## [1] 2 0
## [1] 2 0
## [1] 1 2
## [1] 1 0
## [1] 1 0
## [1] 1 0
## [1] 1 0
## [1] 2 0
## [1] 2 0
## [1] 2 0
## [1] 2 0
## [1] 1 0
## [1] 1 0
## [1] 2 2
## [1] 2 0
```

```
Accidents_Data24$prob.inj <- prob.inj
Accidents_Data24$pred.prob <- ifelse(Accidents_Data24$prob.inj>0.5, "yes", "no")
table(Accidents_Data24$pred.prob)
```

```
##
##  no yes
##  14  10
```

QUESTION-2(3)

#COMPUTING MANUALLY THE NAIVE BAYES CONDITIONAL PROBABILITY OF AN INJURY GIVEN THE WEATHER_R =1 AND TRAF_CON_R =1.

#The Naive Bayes conditional probability is computed using the Naive Bayes formula as follows: #P(INJURY = Yes | WEATHER_R = 1 and TRAF_CON_R = 1) = (P(INJURY = Yes | WEATHER_R = 1) * P(INJURY = Yes | TRAF_CON_R = 1) * P(INJURY = Yes)) / (P(WEATHER_R = 1) * P(TRAF_CON_R = 1))

```
Manual_NB_W1_T1 <- Prob3
cat("Manual Naive Bayes Conditional Probability (Injury = Yes | Weather_R =
1, TRAF_CON_R = 1):", Manual_NB_W1_T1)
```

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

QUESTION-3(4)

#RUNNING A NAIVE BAYES CLASSIFIER ON THE 24 RECORDS AND TWO PREDICTORS. #NOW,WE HAVE TO CHECK THE MODEL OUTPUT TO OBTAIN PROBABILITIES AND CLASSIFCATIONS FOR ALL 24 RECORDS. ##AND THEN, WE ARE COMPARING TO BAYES CLASSIFCATION TO SEE IF THE RESULTING CLASSIFICATIONS ARE EQUIVALENT OR NOT.

```
library(e1071)

NB<-naiveBayes(INJURY ~ ., data = Accidents_Data24)

NBT <- predict(NB, newdata = Accidents_Data24,type = "raw")

Accidents_Data24$nbpred.prob <- NBT[,2] # Transfer the "Yes" nb prediction
library(caret)

NB2 <- train(INJURY ~ TRAF_CON_R + WEATHER_R,
 data = Accidents_Data24, method = "nb")
```

```
## 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_R

## 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_R

## 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_R

## 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_R

## 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_R

## 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_R, WEATHER_R

## 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_R

## 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_R

## 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_R
## 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_R
```

```
## 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_R
```

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

```
predict(NB2, newdata = Accidents_Data24[,c("INJURY", "WEATHER_R", "TRAF_CON_R")])
```

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

```
predict(NB2, newdata = Accidents_Data24[,c("INJURY", "WEATHER_R", "TRAF_CON_R")],
 type = "raw")
```

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

QUESTION-3

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

```
#Splitting the data into training (60%) and validation (40%)

set.seed(123)
TrainIndex <- createDataPartition(Accidents_Data$INJURY, p = 0.6, list =
FALSE)
Train_Data <- Accidents_Data[TrainIndex, ]
Val_Data <- Accidents_Data[-TrainIndex, ]
```

QUESTION-3(1)

Run a naive Bayes classifier on the complete training set with the relevant predictors (and INJURY as the response). Note that all predictors are categorical. Show the confusion matrix.

```
#Splitting the data into training (60%) and validation (40%)

set.seed(123)

trainIndex <- createDataPartition(Accidents_Data$INJURY, p = 0.6, list =
FALSE)
train_data <- Accidents_Data[trainIndex, ]
val_data <- Accidents_Data[-trainIndex, ]

#Creating a naive bayes model with the relavant predictors
nb <- naiveBayes(INJURY ~ WEATHER_R + TRAF_CON_R, data = train_data)

#Predicting on the validation set
```

```r
val_pred <-predict(nb, newdata = val_data)

#Creating a confusion matrix
confusion_matrix <- confusionMatrix(val_pred, val_data$INJURY)
print(confusion_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  1294 1064
##        yes 6994 7520
##
##                Accuracy : 0.5224
##                  95% CI : (0.5148, 0.53)
##     No Information Rate : 0.5088
##     P-Value [Acc > NIR] : 0.0002039
##
##                   Kappa : 0.0326
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.1561
##             Specificity : 0.8760
##          Pos Pred Value : 0.5488
##          Neg Pred Value : 0.5181
##              Prevalence : 0.4912
##          Detection Rate : 0.0767
##    Detection Prevalence : 0.1398
##       Balanced Accuracy : 0.5161
##
##        'Positive' Class : no
##
```

QUESTION-3(2)

```r
#OVERALL ERROR OF THE VALIDATION SET

Overall_Error <- 1 - confusion_matrix$overall["Accuracy"]
cat("overall error of the validation set:", Overall_Error, "\n")
```

```
## overall error of the validation set: 0.477596
```