Main Body of Exercise

Saeid Abolfazli

May 9, 2016

How many parolees are contained in the dataset?

```
parole <- read.table("data/parole.csv",sep=",", header=TRUE)
str(parole)
```

```
## 'data.frame':    675 obs. of  9 variables:
##  $ male             : int  1 0 1 1 1 1 1 0 0 1 ...
##  $ race             : int  1 1 2 1 2 2 1 1 1 2 ...
##  $ age              : num  33.2 39.7 29.5 22.4 21.6 46.7 31 24.6 32.6 29.1 ...
##  $ state            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ time.served      : num  5.5 5.4 5.6 5.7 5.4 6 6 4.8 4.5 4.7 ...
##  $ max.sentence     : int  18 12 12 18 12 18 18 12 13 12 ...
##  $ multiple.offenses: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ crime            : int  4 3 3 1 1 4 3 1 3 2 ...
##  $ violator         : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
table(parole$violator)
```

```
##
##   0   1
## 597  78
```

How many of the parolees in the dataset violated the terms of their parole?

You should be familiar with unordered factors (if not, review the Week 2 homework problem "Reading Test Scores"). Which variables in this dataset are unordered factors with at least three levels? Select all that apply.

```
unique(parole$state)
```

```
## [1] 1 2 3 4
```

In the last subproblem, we identified variables that are unordered factors with at least 3 levels, so we need to convert them to factors for our prediction problem (we introduced this idea in the "Reading Test Scores" problem last week). Using the as.factor() function, convert these variables to factors. Keep in mind that we are not changing the values, just the way R understands them (the values are still numbers).

How does the output of summary() change for a factor variable as compared to a numerical variable?

```
parole$state <- as.factor(parole$state)
parole$crime <- as.factor(parole$crime)
summary(parole$crime)
```

```
##   1   2   3   4
## 315 106 153 101
```

```
table(parole$crime)
```

```
##
##   1   2   3   4
## 315 106 153 101
```

To ensure consistent training/testing set splits, run the following 5 lines of code (do not include the line numbers at the beginning):

```
set.seed(144)
library(lubripack)
lubripack("caTools")
```

```
##
## Bellow Packages Successfully Installed:
##
## caTools
##    TRUE
```

```
index <- sample.split(parole$violator,0.7)
paroleTrain <- parole[index,]
paroleTest <- parole[!index,]
```

If you tested other training/testing set splits in the previous section, please re-run the original 5 lines of code to obtain the original split.

Using glm (and remembering the parameter family="binomial"), train a logistic regression model on the training set. Your dependent variable is "violator", and you should use all of the other variables as independent variables.

```
Model1 <- glm(violator~.,data=paroleTrain,family=binomial)
summary(Model1)
```

```
##
## Call:
## glm(formula = violator ~ ., family = binomial, data = paroleTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7041  -0.4236  -0.2719  -0.1690   2.8375
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.2411574  1.2938852  -3.278  0.00105 **
## male               0.3869904  0.4379613   0.884  0.37690
## race               0.8867192  0.3950660   2.244  0.02480 *
## age               -0.0001756  0.0160852  -0.011  0.99129
## state2             0.4433007  0.4816619   0.920  0.35739
## state3             0.8349797  0.5562704   1.501  0.13335
## state4            -3.3967878  0.6115860  -5.554 2.79e-08 ***
## time.served       -0.1238867  0.1204230  -1.029  0.30359
```

```
## max.sentence       0.0802954  0.0553747   1.450  0.14705
## multiple.offenses  1.6119919  0.3853050   4.184 2.87e-05 ***
## crime2             0.6837143  0.5003550   1.366  0.17180
## crime3            -0.2781054  0.4328356  -0.643  0.52054
## crime4            -0.0117627  0.5713035  -0.021  0.98357
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 340.04  on 472  degrees of freedom
## Residual deviance: 251.48  on 460  degrees of freedom
## AIC: 277.48
##
## Number of Fisher Scoring iterations: 6
```

Consider a parolee who is male, of white race, aged 50 years at prison release, from the state of Maryland, served 3 months, had a maximum sentence of 12 months, did not commit multiple offenses, and committed a larceny. Answer the following questions based on the model's predictions for this individual.

According to the model, what are the odds this individual is a violator?

```
Logit <-  -4.2411573912 + 0.3869904022 + 0.8867192411 - (50*0.0001756156) +  (0*0.4433006515) + (0*0.834

Odds <- exp(Logit)
Odds
```

```
## [1] 0.1825685
```

According to the model, what is the probability this individual is a violator?

```
probability <- 1/(1+exp(-Logit))
probability
```

```
## [1] 0.154383
```

Obtain the model's predicted probabilities for parolees in the testing set, remembering to pass type="response".

What is the maximum predicted probability of a violation?

```
predViolation <- predict(Model1,type="response",newdata = paroleTest)
max(predViolation)
```

```
## [1] 0.9072791
```

What is the model's sensitivity?

```
table(paroleTest$violator ,as.numeric(predViolation >= 0.5))
```

```
##
##       0   1
##   0 167  12
##   1  11  12
```

Answer: 0.5217

What is the model's specificity?

Answer:0.933

What is the model's accuracy?

Answer: 0.886

What is the accuracy of a simple model that predicts that every parolee is a non-violator?

```
table(parole$violator)
```

```
##
##   0   1
## 597  78
```

Answer : 597/(597+78) = 0.88

Consider a parole board using the model to predict whether parolees will be violators or not. The job of a parole board is to make sure that a prisoner is ready to be released into free society, and therefore parole boards tend to be particularily concerned about releasing prisoners who will violate their parole. Which of the following most likely describes their preferences and best course of action?

- The board assigns more cost to a false negative than a false positive, and should therefore use a logistic regression cutoff higher than 0.5.
- The board assigns more cost to a false negative than a false positive, and should therefore use a logistic regression cutoff less than 0.5.
- The board assigns equal cost to a false positive and a false negative, and should therefore use a logistic regression cutoff equal to 0.5.
- The board assigns more cost to a false positive than a false negative, and should therefore use a logistic regression cutoff higher than 0.5.
- The board assigns more cost to a false positive than a false negative, and should therefore use a logistic regression cutoff less than 0.5.

Answer: 2

Which of the following is the most accurate assessment of the value of the logistic regression model with a cutoff 0.5 to a parole board, based on the model's accuracy as compared to the simple baseline model?

- The model is of limited value to the board because it cannot outperform a simple baseline, and using a different logistic regression cutoff is unlikely to improve the model's value.
- The model is of limited value to the board because it cannot outperform a simple baseline, and using a different logistic regression cutoff is likely to improve the model's value.
- The model is likely of value to the board, and using a different logistic regression cutoff is unlikely to improve the model's value.
- The model is likely of value to the board, and using a different logistic regression cutoff is likely to improve the model's value.

Using the ROCR package, what is the AUC value for the model?

```
lubripack("ROCR")
```

```
## Warning: package 'gplots' was built under R version 3.2.4
```

```
##
## Attaching package: 'gplots'
##
## The following object is masked from 'package:stats':
##
##     lowess
```

```
##
## Bellow Packages Successfully Installed:
##
## ROCR
## TRUE
```

```
ROCRPred <- prediction(predViolation, paroleTest$violator)
perfMode <- performance(ROCRPred,"auc")@y.values[[1]]
```

Describe the meaning of AUC in this context.

- The probability the model can correctly differentiate between a randomly selected parole violator and a randomly selected parole non-violator.
- The probability the model can correctly differentiate between a randomly selected parole violator and a randomly selected parole non-violator. - correct The model's accuracy at logistic regression cutoff 0.5.
- The model's accuracy at the logistic regression cutoff at which it is most accurate.

Answer: 1

Identifying Bias in Observational Data

Our goal has been to predict the outcome of a parole decision, and we used a publicly available dataset of parole releases for predictions. In this final problem, we'll evaluate a potential source of bias associated with our analysis. It is always important to evaluate a dataset for possible sources of bias.

The dataset contains all individuals released from parole in 2004, either due to completing their parole term or violating the terms of their parole. However, it does not contain parolees who neither violated their parole nor completed their term in 2004, causing non-violators to be underrepresented. This is called "selection bias" or "selecting on the dependent variable," because only a subset of all relevant parolees were included in our analysis, based on our dependent variable in this analysis (parole violation). How could we improve our dataset to best address selection bias?