PREDICTING PAROLE VIOLATORS

In many criminal justice systems around the world, inmates deemed not to be a threat to society are released from prison under the parole system prior to completing their sentence. They are still considered to be serving their sentence while on parole, and they can be returned to prison if they violate the terms of their parole.

Parole boards are charged with identifying which inmates are good candidates for release on parole. They seek to release inmates who will not commit additional crimes after release. In this problem, we will build and validate a model that predicts if an inmate will violate the terms of his or her parole. Such a model could be useful to a parole board when deciding to approve or deny an application for parole.

For this prediction task, we will use data from the United States 2004 National Corrections Reporting Program (http://www.icpsr.umich.edu/icpsrweb/NACJD/series/38/studies/26521?archive=NACJD&sortBy=7), a nationwide census of parole releases that occurred during 2004. We limited our focus to parolees who served no more than 6 months in prison and whose maximum sentence for all charges did not exceed 18 months. The dataset contains all such parolees who either successfully completed the term of parole during 2004 or those who violated the terms of their parole during that year. The dataset contains the following variables:

- **male**: 1 if the parolee is male, 0 if female
- **race**: 1 if the parolee is white, 2 otherwise
- **age**: the parolee's age in years at release from prison
- **state**: a code for the parolee's state. 2 is Kentucky, 3 is Louisiana, 4 is Virginia, and 1 is any other state. The three states were selected due to having a high representation in the dataset.
- **time.served**: the number of months the parolee served in prison (limited by the inclusion criteria to not exceed 6 months).
- **max.sentence**: the maximum sentence length for all charges, in months (limited by the inclusion criteria to not exceed 18 months).
- **multiple.offenses**: 1 if the parolee was incarcerated for multiple offenses, 0 otherwise.
- **crime**: a code for the parolee's main crime leading to incarceration. 2 is larceny, 3 is drug-related crime, 4 is driving-related crime, and 1 is any other crime.
- **violator**: 1 if the parolee violated the parole, and 0 if the parolee completed the parole without violation.

PROBLEM 1.1 - LOADING THE DATASET (1 point possible)

Load the dataset parole.csv (/c4x/MITx/15.071x/asset/parole.csv) into a data frame called parole, and investigate it using the str() and summary() functions.

How many parolees are contained in the dataset?

Show Answer You have used 0 of 3 submissions

PROBLEM 1.2 - LOADING THE DATASET (1 point possible)

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

Show Answer You have used 0 of 3 submissions

PROBLEM 1.3 - LOADING THE DATASET (1 point possible)

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?

- ☐ male
- ☐ race
- ☐ age
- ☐ state
- ☐ time.served
- ☐ max.sentence
- ☐ multiple.offenses
- ☐ crime
- ☐ violator

Show Answer You have used 0 of 2 submissions

PROBLEM 2.1 - PREPARING THE DATASET (1 point possible)

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?

- ○ The output becomes similar to that of the table() function applied to that variable
- ○ The output becomes similar to that of the str() function applied to that variable
- ○ There is no change

Show Answer You have used 0 of 1 submissions

PROBLEM 2.2 - PREPARING THE DATASET (1 point possible)

Why are we taking this step of preparing the variables before splitting the data into a training and testing set?

- ○ After the training/testing set split, we would be unable to take these preparatory steps.
- ○ Preparing the data before splitting the dataset saves work: we only need to do these steps once instead of twice

- ○ There is no difference between preparing the dataset before or after the training/testing set split.

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.1 - SPLITTING INTO A TRAINING AND TESTING SET (1 point possible)

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

1) set.seed(144)

2) library(caTools)

3) split = sample.split(parole$violator, SplitRatio = 0.7)

4) train = subset(parole, split == TRUE)

5) test = subset(parole, split == FALSE)

Roughly what proportion of parolees have been allocated to the training and testing sets?

- ○ 70% to the training set, 30% to the testing set
- ○ 50% to the training set, 50% to the testing set
- ○ 30% to the training set, 70% to the testing set

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.2 - SPLITTING INTO A TRAINING AND TESTING SET (3 points possible)

Now, suppose you re-ran lines [1]-[5] of Problem 3.1. What would you expect?

- ○ The exact same training/testing set split as the first execution of [1]-[5]
- ○ A different training/testing set split from the first execution of [1]-[5]

If you instead ONLY re-ran lines [3]-[5], what would you expect?

- ○ The exact same training/testing set split as the first execution of [1]-[5]
- ○ A different training/testing set split from the first execution of [1]-[5]

If you instead called set.seed() with a different number and then re-ran lines [3]-[5] of Problem 3.1, what would you expect?

- ○ The exact same training/testing set split as the first execution of [1]-[5]
- ○ A different training/testing set split from the first execution of [1]-[5]

Show Answer *You have used 0 of 1 submissions*

PROBLEM 4.1 - BUILDING A LOGISTIC REGRESSION MODEL (1 point possible)

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.

What variables are significant in this model? Significant variables should have a least one star, or should have a probability less than 0.05 (the column Pr(>|z|) in the summary output).

- ☐ male
- ☐ race
- ☐ age
- ☐ state2
- ☐ state3
- ☐ state4
- ☐ time.served
- ☐ max.sentence
- ☐ multiple.offenses
- ☐ crime2
- ☐ crime3
- ☐ crime4

Show Answer *You have used 0 of 3 submissions*

PROBLEM 4.2 - BUILDING A LOGISTIC REGRESSION MODEL (1 point possible)

What can we say based on the coefficient of the multiple.offenses variable?

The following two properties might be useful to you when answering this question:

1) If we have a coefficient c for a variable, then that means the log odds (or Logit) are increased by c for a unit increase in the variable.

2) If we have a coefficient c for a variable, then that means the odds are multiplied by e^c for a unit increase in the variable.

- ○ Our model predicts that parolees who committed multiple offenses have 1.61 times higher odds of being a violator than the average parolee.
- ○ Our model predicts that a parolee who committed multiple offenses has 1.61 times higher odds of being a violator than a parolee who did not commit multiple offenses but is otherwise identical.
- ○ Our model predicts that parolees who committed multiple offenses have 5.01 times higher odds of being a violator than the average parolee.
- ○ Our model predicts that a parolee who committed multiple offenses has 5.01 times higher odds of being a violator than a parolee who did not commit multiple offenses but is otherwise identical.

Show Answer *You have used 0 of 2 submissions*

PROBLEM 4.3 - BUILDING A LOGISTIC REGRESSION MODEL (2 points possible)

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. (HINT: You should use the coefficients of your model, the Logistic Response Function, and the Odds equation to solve this problem.)

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

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

Show Answer *You have used 0 of 5 submissions*

PROBLEM 5.1 - EVALUATING THE MODEL ON THE TESTING SET (1 point possible)

Use the predict() function to 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?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 5.2 - EVALUATING THE MODEL ON THE TESTING SET (3 points possible)

In the following questions, evaluate the model's predictions on the test set using a threshold of 0.5.

What is the model's sensitivity?

What is the model's specificity?

What is the model's accuracy?

Show Answer *You have used 0 of 5 submissions*

PROBLEM 5.3 - EVALUATING THE MODEL ON THE TESTING SET (1 point possible)

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

Show Answer *You have used 0 of 3 submissions*

PROBLEM 5.4 - EVALUATING THE MODEL ON THE TESTING SET (1 point possible)

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 with 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.

Show Answer *You have used 0 of 2 submissions*

PROBLEM 5.5 - EVALUATING THE MODEL ON THE TESTING SET (1 point possible)

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.

Show Answer *You have used 0 of 1 submissions*

PROBLEM 5.6 - EVALUATING THE MODEL ON THE TESTING SET (1 point possible)

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

Show Answer You have used 0 of 3 submissions

PROBLEM 5.7 - EVALUATING THE MODEL ON THE TESTING SET (1 point possible)

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 model's accuracy at logistic regression cutoff 0.5.
- ○ The model's accuracy at the logistic regression cutoff at which it is most accurate.

Show Answer You have used 0 of 1 submissions

PROBLEM 6.1 - IDENTIFYING BIAS IN OBSERVATIONAL DATA (1 point possible)

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?

- ○ There is no way to address this form of biasing.
- ○ We should use the current dataset, expanded to include the missing parolees. Each added parolee should be labeled with violator=0, because they have not yet had a violation.
- ○ We should use the current dataset, expanded to include the missing parolees. Each added parolee should be labeled with violator=NA, because the true outcome has not been observed for these individuals.
- ○ We should use a dataset tracking a group of parolees from the start of their parole until either they violated parole or they completed their term.

Show Answer You have used 0 of 1 submissions