University of Waterloo

Department of Management Engineering

MSCI 433 – Applications of Management Engineering

Assignment 1: Predicting Parole Violators

Sarbajoy Majumdar (20531951)

Part 1: Descriptive Analysis

1.1: Summary of Training Data

After the parole data was split into two sets, the training set (stored in variable named “train” in R) and the testing set (stored in variable named “test” in R), the general data of the training data was gathered.

Table 1 showcases that there are 378 male parolees in the data set and 95 female parolees in the data set.

Table 1: Gender Statistics of Parolee Training Set

Gender	Count
Male	378
Female	95

Table 2 showcases that there are 269 white parolees in the data set and 204 non-white parolees in the data set.

Table 2: Race Statistics of Parolee Training Set

Race	Count
White	269
Non-White	204

Table 3 showcases the age statistics of the parolees. It is observed that the median age and the mean age are fairly close to each other.

Table 3: Age Statistics of Parolee Training Set in years

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
18.4	25.8	34.9	35.0	43.0	43.0

Table 4 showcases the states breakdown of all the parolee in the data set. It is observed that Virginia the state with the highest number of parolees.

Table 4: State Statistics of Parolee Training Set

State	Count
Kentucky	78
Louisiana	56
Virginia	237
Other	102

Table 5 showcases the time served statistics of the parolees.

Table 5: Time Served Statistics of Parolee Training Set in months

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
0.0	3.2	4.4	4.2	5.1	6.0

Table 6 showcases the maximum sentence statistics of the parolees.

Table 6: Time Served Statistics of Parolee Training Set in months

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
1.0	12.0	12.0	12.9	15.0	18.0

Table 7 showcases that there are 270 multiple offenders in the data set and 203 single offenders in the data set.

Table 2: Multiple Offender Statistics of Parolee Training Set

Gender	Count
Multiple Offenders	270
Single Offenders	203

Table 8 showcases the crime breakdown of all the parolee in the data set. It is observed that drugs is the crime with the highest number of parolees.

Table 8: Crime Statistics of Parolee Training Set

Crime	Count
Driving	70
Drugs	105
Larceny	72
Other	226

Table 9 showcases that there are 270 multiple offenders in the data set and 203 single offenders in the data set.

Table 2: Violator Statistics of Parolee Training Set

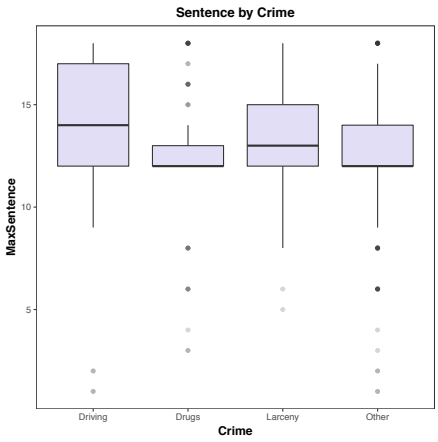
Parole Violator Status	Count
Violator	55
Non-Violator	418

1.2 Data Insights

With the general statistics considered, the next step is to derive insights on given data.

The first graph showcases the average sentence for each type of crime. It is seen that whilst drug-related offenses have a lower average sentencing time compared to other crimes and has a smaller range of sentencing times, there are numerous outliers for drug-related offenses. This signifies that drug sentences might not have consistently defined and hence it could be inferred that drug offences might not be a significant variable to determine who would violate parole.

Figure 1: Boxplot of Maximum Sentence by Crime

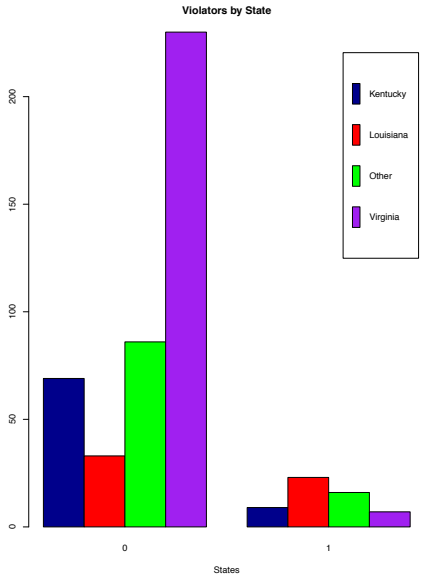


Seeing the range of maximum sentences per crime is insightful in discovering the length of the sentence for each type of crime and this information could be insightful in determining the validity of parole for some crime over others.

For the next few graph plots, the statistics for violators were graphed based off of four different metrics: state, crime, race and gender. Since the overall goal of the assignment is to predict the validity of a parole application, it was deemed appropriate to compare parole violators with these four metrics.

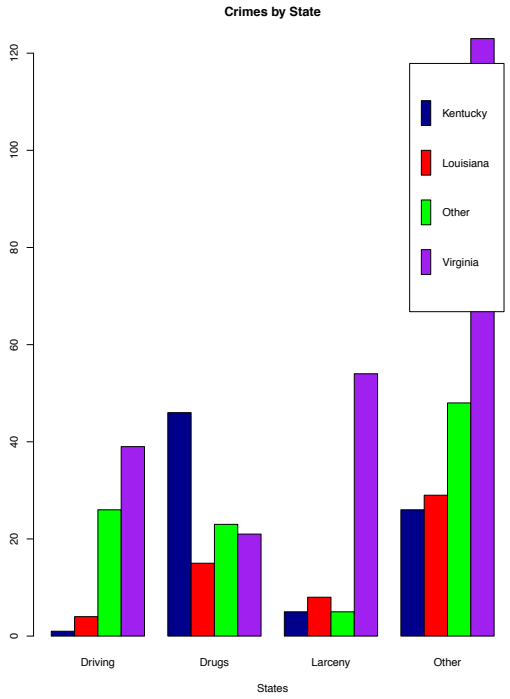
The first metric compared was parole violators to the state they are incarcerated as seen in Figure 2. It is seen that Louisiana is the state with the most parole violators. This could indicate that Louisiana would have a higher likelihood of parole violators compared to other states.

Figure 2: Histogram of Violator by State



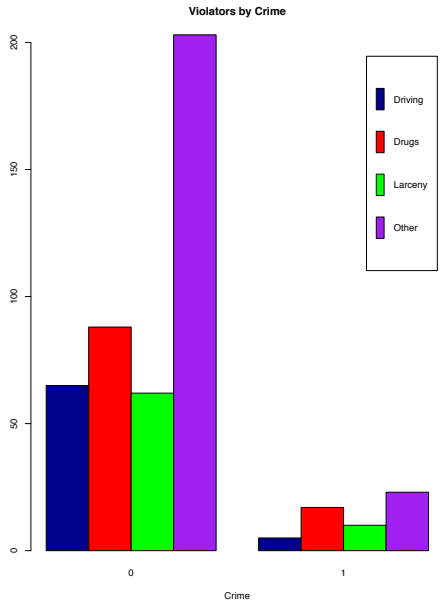
After gathering that data on states to violators, the next relationship checked was between crimes committed and the state. A unique insight drawn from this data is that Kentucky has the highest rate of drug crimes eligible for parole compared to any other states. An interpretation drawn is that since Kentucky has one of the lowest numbers of parole violators yet the highest number of drug offenders, there might be a chance that being from Kentucky and/or committing a drug crime might not be a significant variable to determine who would violate parole.

Figure 3: Histogram of Crime by State



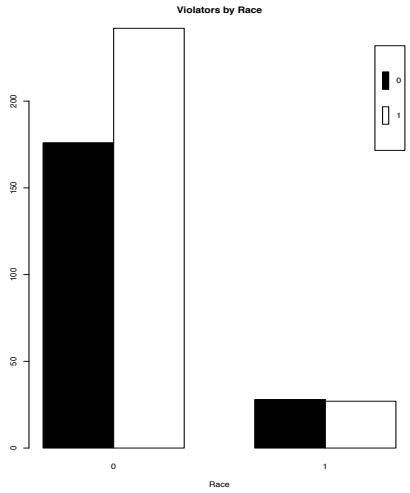
Next the relationship between crimes committed and violators were checked. It was observed that drug crimes have the highest number of parole violators. This fact contradicts with the insight generated by the previous data visualization which hypothesized that drug crime might not play a role in determining parole violators. Thus, it would be up to the logistic regression model to determine whether drug crime is a significant factor to determine parole violators.

Figure 4: Histogram of Violator by Crime



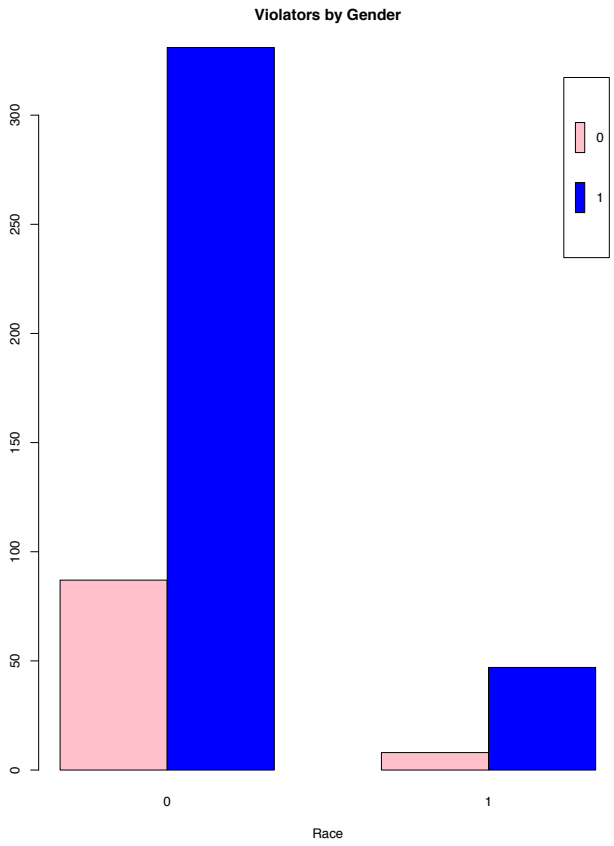
Next, the comparison of race and violators was compared. It was observed that non-whites form a slight majority of the parole violators over the whites. It could be interpreted that race would be a significant factor in determining a parole violator.

Figure 5: Histogram of Violator by Race



The last data that was checked was the gender and violators. Males did form the majority of the violators, but males also formed the majority of the population. Thus, it would be hard to conclude whether gender plays a role in determining parole violators.

Figure 6: Histogram of Violator by Gender



Part 2: Predictive Analysis

2.1: Logistic Regression

Figure 7: Summary of Logistic Regression

```
Call:
glm(formula = Violator ~ Male + RaceWhite + Age + State + TimeServed +
    MaxSentence + MultipleOffenses + Crime, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5641  -0.4253  -0.2650  -0.1709   2.8759

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.791293   1.444104  -1.933   0.0532 .
Male              0.690480   0.478937   1.442   0.1494
RaceWhite        -0.700040   0.402230  -1.740   0.0818 .
Age               0.007743   0.016557   0.468   0.6400
StateLouisiana    0.242354   0.641694   0.378   0.7057
StateOther       -0.074029   0.524126  -0.141   0.8877
StateVirginia    -3.462410   0.723268  -4.787 1.69e-06 ***
TimeServed       -0.132017   0.116575  -1.132   0.2574
MaxSentence       0.046429   0.050907   0.912   0.3617
MultipleOffenses  1.818978   0.418170   4.350 1.36e-05 ***
CrimeDrugs        0.215612   0.646076   0.334   0.7386
CrimeLarceny      1.261845   0.684576   1.843   0.0653 .
CrimeOther       -0.078874   0.587515  -0.134   0.8932
---
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: 248.00  on 460  degrees of freedom
AIC: 274

Number of Fisher Scoring iterations: 6
```

Based on Figure 7, the median of the deviance residuals of the logistic regression model is -0.27. The Fisher Scoring iterations of the model is 6, and the AIC is 274.

Based on the logistic regression model, the variables of StateVirginia and MultipleOffenses has the most significance in determining who would violate parole with a significance code of 0.001. RaceWhite and CrimeLarceny are also significant with a significance code of 0.1.

The data of a prisoner in the assignment, who is assigned to a variable labelled prisonMike in the R code, has been inserted into the logistic regression model via a predict function. Based off of that, it is determined that there is a 25.6% chance that the prisoner in question would violate their parole.

2.2: Confusion Matrix

Figure 8: Confusion Matrix of Logistic Regression Model on Test Data

	FALSE	TRUE
0	174	5
1	17	6

The confusion matrix is shown in Figure 8. The TRUE column indicates that the model predicts the convict to be a parole violator and the FALSE column indicates that the model predicts the convict to not be a parole violator. The row 0 indicates that the parolee was actually not a violator and 1 indicates that the parolee was actually a violator.

The number of false positives in the model is 5 and the number of false negatives is 17. The false positive rate is 0.0279. The false negative rate is 0.739. The overall accuracy of the model is 0.891.

The baseline model has an accuracy of 0.886. Based off of accuracy alone, our model performs better than the baseline. Based on false positives, the baseline does better because the baseline does not produce a false positive rate. The false negative rate of the baseline is 1, which is significantly higher than the false negative rate of the model.

2.3: Parole Board's Requirements

The parole board would be more concerned with the false negative errors because a false negative would mean that the model indicates that the criminal would not violate parole but would actually violate parole. Violating their parole could possibly mean a potential repeat of the crimes that the criminals have done, and hence that could endanger the lives of people in society.

The parole board should adjust their threshold to be lower than 0.5 to reduce false negative errors. However, if they adjust their threshold to be exactly 0, it would indicate that nobody gets parole and hence that would eliminate the need for having a parole board (or even parole) to begin with.

2.4: Area Under Curve Analysis

The Area Under the Curve (AUC) is calculated to be 0.815. The ideal AUC would be 1, which would indicate that the model is ideal in terms of predicting parole violators. Thus, the closer the AUC value of the model is to 1, the more accurate the model is in determining whether parolees would violate or not. Thus, the current model is fairly accurate in determining who would be a parole violator.

Considering the accuracy of the baseline model and further threshold adjustments, it is seen that the threshold at which the baseline model and the logistic regression model will have the same accuracy is when the threshold is higher, thus it is recommended for the parole board to move onto the newly developed model and not use the baseline model. The model would be accurate most of the times and would actually determine some of the parolees to be parole violators and thus preventing them from getting parole, whereas the older model would grant parole to everyone regardless of their background and eligibility.

However, if the parole board does want to reduce the false negative rate, they would have to lower the threshold of the model which in turn would result in a lower accuracy of the model.

Resources

https://github.com/burun/MITx_AnalyticsEdge_2015/blob/master/Unit3/Assignments/Predicting%20Parole%20Violators.R#L89

http://gim.unmc.edu/dxtests/roc3.htm

Appendix A: R Code

```
parole= read.csv("Downloads/Parole.csv")
library(caTools)
set.seed(1951)
parole$split = sample.split(parole$Violator,SplitRatio=0.7)
train = subset(parole, split==TRUE)
test = subset(parole, split==FALSE)

## question1
nrow(train)
table(train$Male)
table(train$RaceWhite)
table(train$State)
table(train$MultipleOffenses)
table(train$Crime)
table(train$Violator)
summary(train)

##question2
library(ggplot2)
str(train)
plotCrimeMaxSentence = ggplot(train, aes(x=Crime, y=MaxSentence)) + geom_boxplot(fill="slateblue", alpha=0.2) +ggtitle("Sentence by Crime") + theme_bw() + theme(plot.title=element_text(hjust=0.5, face="bold"), axis.title=element_text(size=12,color="black",face="bold"),panel.grid.major=element_blank(),panel.grid.minor=element_blank())
ggsave("CrimeSentence.pdf",plotCrimeMaxSentence,width=6,height=6, units="in")
countStateViolator <- table(train$State, train$Violator)
```

```
barplot(countStateViolator, main="Violators by State",xlab="States", col=c("darkblue","red","green","purple"),legend = rownames(countStateViolator), beside=TRUE)
countStateCrime <- table(train$State, train$Crime)
barplot(countStateCrime, main="Crimes by State",xlab="States", col=c("darkblue","red","green","purple"),legend = rownames(countStateCrime), beside=TRUE)
countCrimeViolator <- table(train$Crime, train$Violator)
barplot(countCrimeViolator, main="Violators by Crime",xlab="Crime", col=c("darkblue","red","green","purple"),legend = rownames(countCrimeViolator), beside=TRUE)
countRaceViolator <- table(train$RaceWhite, train$Violator)
barplot(countRaceViolator, main="Violators by Race",xlab="Race", col=c("black","white"),legend = rownames(countRaceViolator), beside=TRUE)
countGenderViolator <- table(train$Male, train$Violator)
barplot(countGenderViolator, main="Violators by Gender",xlab="Race", col=c("pink","blue"),legend = rownames(countGenderViolator), beside=TRUE)


##Log Regression
ParoleViolator=glm(Violator~Male+RaceWhite+Age+State+TimeServed+MaxSentence+MultipleOffenses+Crime, data=train,family=binomial)
summary(ParoleViolator)

##predicting prisoner's chance of violating parole
prisonMike=data.frame(Male=1,RaceWhite=1,Age=50,State="Other",TimeServed=3.0, MaxSentence=12, MultipleOffenses=0,Crime="Larceny")
predict(ParoleViolator, prisonMike,type="response")
```

```
##Confusion matrix
predictTest = predict(ParoleViolator, type = "response", newdata = test)
table(test$Violator, predictTest>0.5)
5/(5+174) #false positive
17/(17+6) #false negative
(174+6)/(174+5+6+17) #overall accuracy

##Baseline Confusion
table(test$Violator)
179/(179+23)

#AUC
install.packages("ROCR")
library(ROCR)
ROCRpred = prediction(predictTest, test$Violator)
as.numeric(performance(ROCRpred, "auc")@y.values)
```