Assignment 4: Logistic Regression for Parole Reform

Zhanchao Yang

2025-04-04

Scenario

The Governor of Georgia wants to replace subjective parole decisions with a consistent, data-driven policy. You've been hired to develop a logistic regression model to predict the risk of recidivism — and recommend a cutoff value that will be adopted statewide.

Sensitivity vs. Specificity

- Sensitivity: Sensitivity represents the proportion of actual positive cases that the model correctly identifies as positive. A high sensitivity indicates the model has a low rate of false negatives, meaning it more accurately identifies individuals who belong to the positive case. In this specific scenario, sensitivity measures the percentage of recidivists that are correctly predicted as recidivists.In other words, the model detects well on individuals who are likely to re-offend.
- Specificity: Specificity measures the proportion of actual negative cases that the model correctly predicts as negative. A high specificity indicates the model has a low rate of false positives, meaning identifies individuals who do not belong to the positive case. In this specific scenario, specificity measures the percentage of non-recidivists that are correctly predicted as non-recidivists. In other words, the model detects well on individuals who are unlikely to re-offend.

In my opinion, sensitivity should be prioritized over specificity in this scenario. In parole reform, the primary goal is to reduce recidivism and ensure that individuals released from prison do not pose a threat to public safety. Prioritizing sensitivity means that we take extra caution by keeping those who are likely to re-offend in custody, even if it detains some individuals who might not actually commit another crime. The government could then offer compensation to those later proven innocent after the jury and trial.

In a different scenario, specificity may be more important than the sensitivity. For example, some crucial resources like prison are limited or government is losing trust from the public as too many innocent people was detained. In this case, the government may want to prioritize specificity to ensure that individuals who are not likely to re-offend are not unnecessarily detained. This would help maintain public trust and ensure that resources are allocated efficiently.

Data exploration and Data cleaning

Data cleaning

```
recidivism_ga <- data %>%
  mutate(recidivism_yes = if_else(Recidivism_Within_3years == "true", 1, 0))

recidivism_ga <- na.omit(recidivism_ga)
```

Training and Testing Partition

```
set.seed(1234)
trainIndex <- createDataPartition(as.factor(recidivism_ga$recidivism_yes), p = 0.7, list = FALSE)
train <- recidivism_ga[trainIndex, ]
test <- recidivism_ga[-trainIndex, ]
```
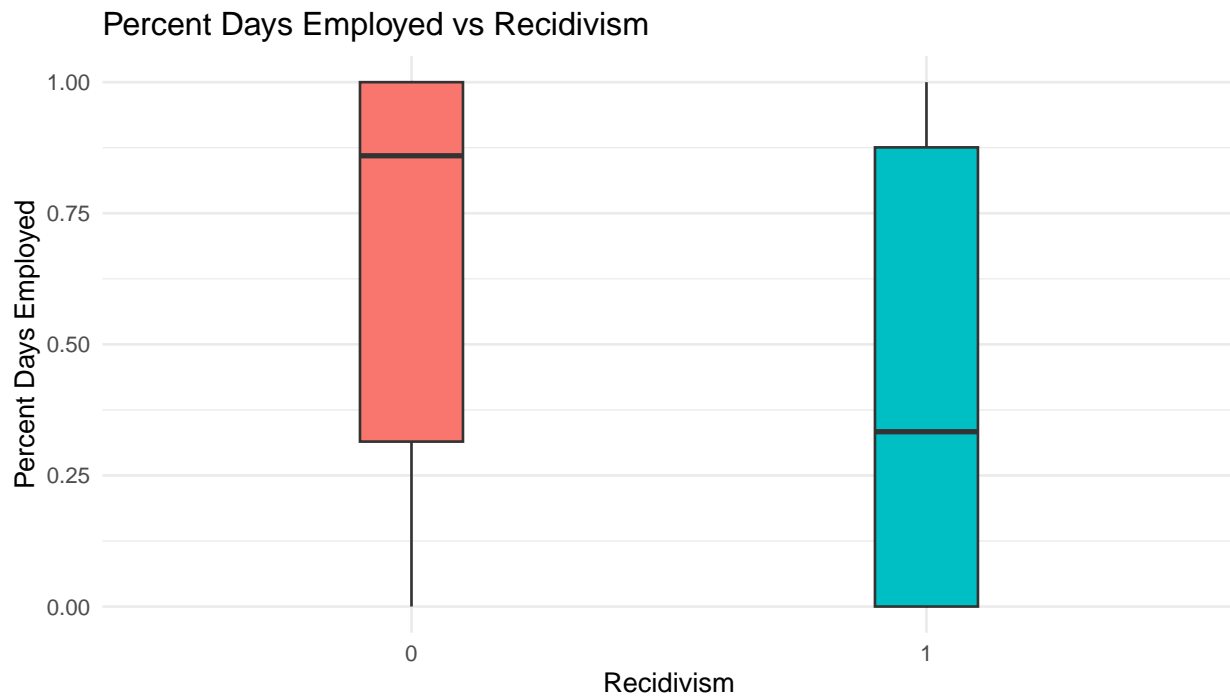
Key predictors

- **Age at Release**: Age at release is a significant predictor of recidivism. Younger individuals are more likely to re-offend compared to older individuals.
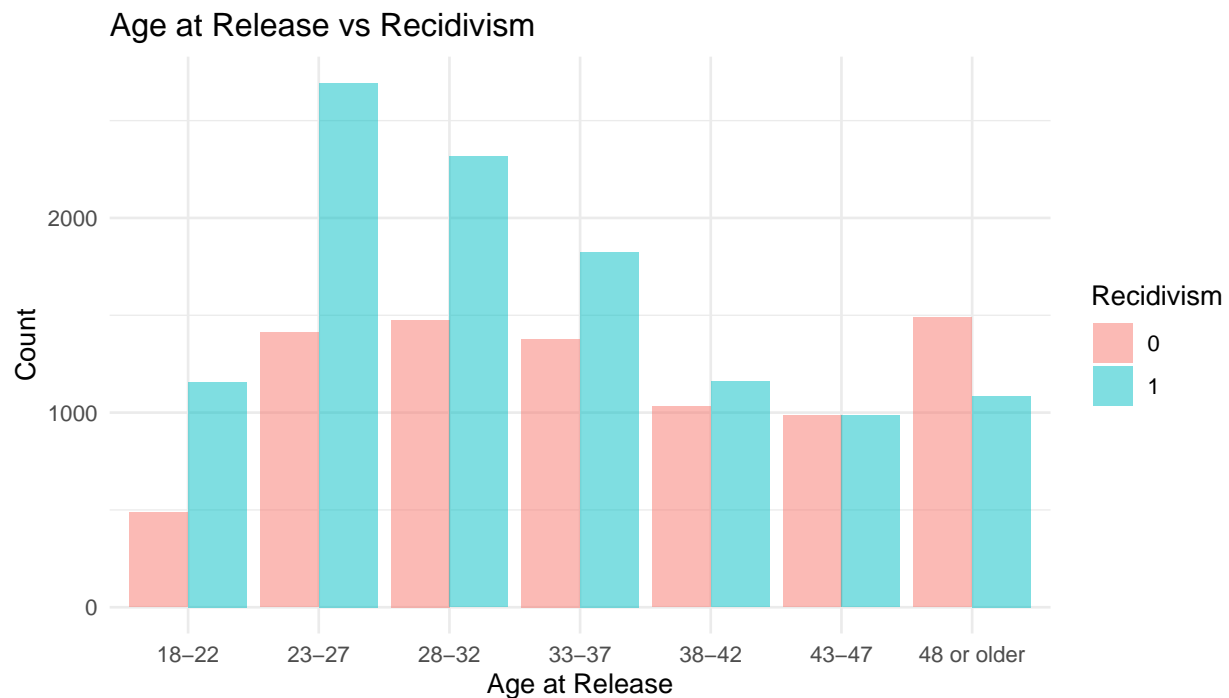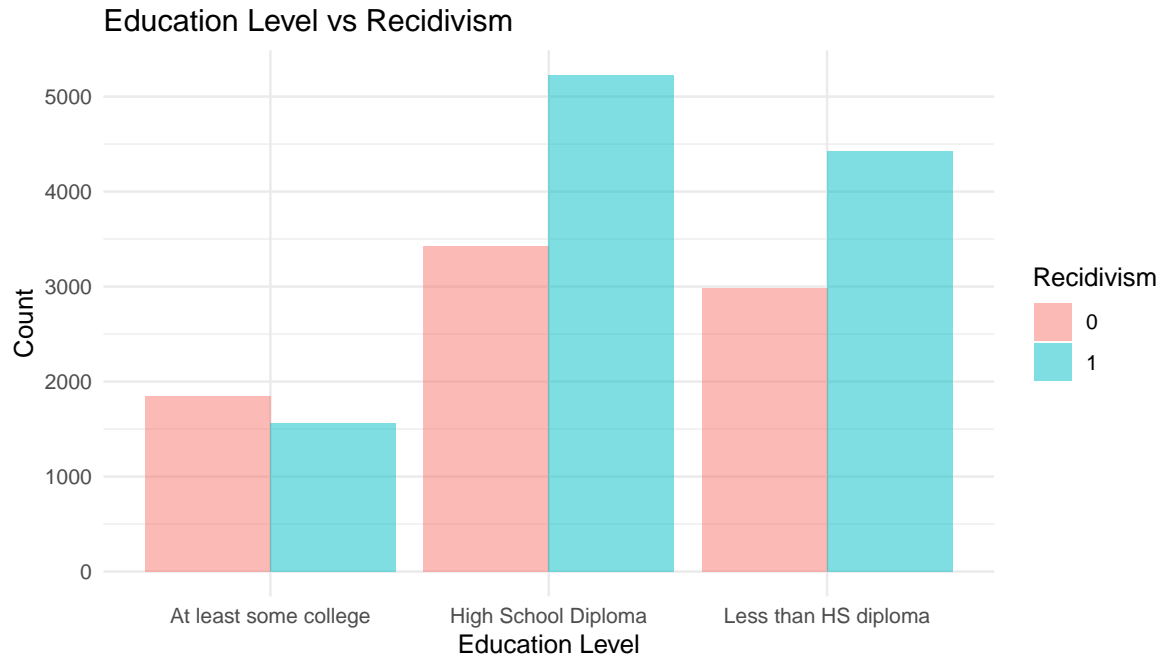

- **Gang Affiliated**: Gang affiliation is a significant predictor of recidivism. Individuals who are affiliated with gangs are more likely to re-offend compared to those who are not.

Gang Affiliated vs Recidivism
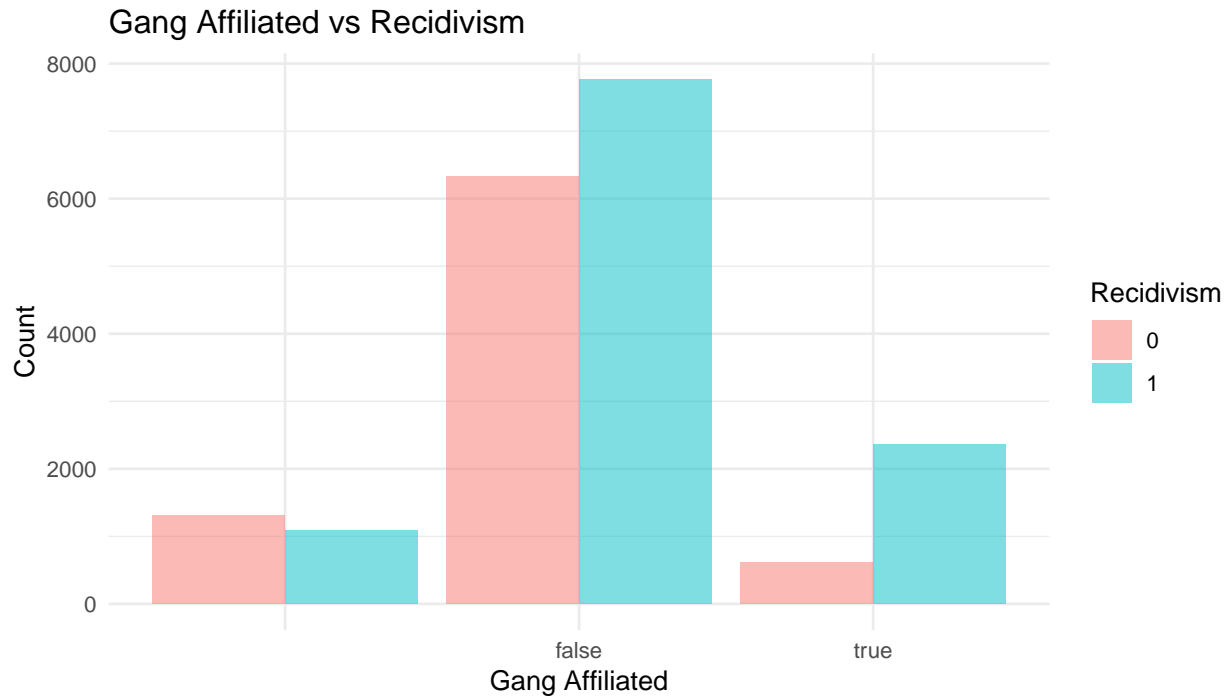
- **Percent Days Employed**: Percent days employed is a significant predictor of recidivism. Individuals who are employed for a higher percentage of days are less likely to re-offend compared to those who are unemployed.

Percent Days Employed vs Recidivism

- **Education Level**: Education level is a significant predictor of recidivism. Individuals with higher education levels are less likely to re-offend compared to those with lower education levels.

Education Level vs Recidivism

- **Prison Offense**: Prison offense is a significant predictor of recidivism. Individuals with more serious offenses like murder are more likely to re-offend compared to those with less serious offenses.

Education Level vs Recidivism
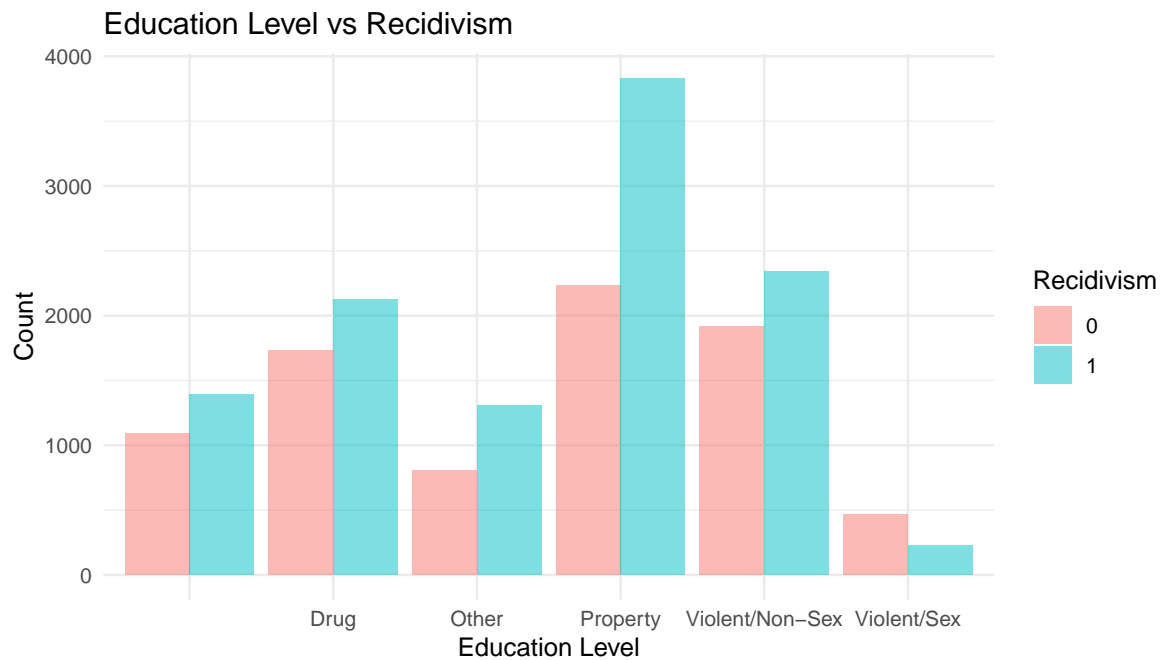
Final Model

After testing several key predictors, I got the final model as following:

```
model <- glm(recidivism_yes ~ Age_at_Release + Gang_Affiliated + Percent_Days_Employed
              + Prison_Offense + Education_Level ,
              data = train, family = "binomial")
```

```r
coeff_table <- summary(model)$coefficients

# Calculate the Odds Ratios by exponentiating the coefficient estimates
odds_ratio <- exp(coeff_table[, "Estimate"])

# Append the odds_ratio as a new column to your coefficients table
coeff_table <- cbind(coeff_table, Odds_Ratio = odds_ratio)

kable(coeff_table,
      caption = "Logistic Regression Coefficient Summary",
      digits = 3)
```

Table 1: Logistic Regression Coefficient Summary

	Estimate	Std. Error	z value	Pr(>|z|)	Odds_Ratio
(Intercept)	0.481	0.108	4.470	0.000	1.617
Age_at_Release23-27	-0.032	0.081	-0.400	0.689	0.968
Age_at_Release28-32	-0.100	0.082	-1.222	0.222	0.905
Age_at_Release33-37	-0.169	0.084	-2.006	0.045	0.845
Age_at_Release38-42	-0.296	0.089	-3.331	0.001	0.744
Age_at_Release43-47	-0.427	0.091	-4.669	0.000	0.652
Age_at_Release48 or older	-0.783	0.087	-8.966	0.000	0.457
Gang_Affiliatedfalse	0.464	0.056	8.271	0.000	1.590
Gang_Affiliatedtrue	1.318	0.078	16.918	0.000	3.735
Percent_Days_Employed	-1.375	0.046	-29.579	0.000	0.253
Prison_OffenseDrug	-0.019	0.066	-0.279	0.780	0.982
Prison_OffenseOther	0.196	0.077	2.543	0.011	1.217
Prison_OffenseProperty	0.339	0.062	5.427	0.000	1.403
Prison_OffenseViolent/Non-Sex	-0.111	0.066	-1.684	0.092	0.895
Prison_OffenseViolent/Sex	-0.932	0.116	-8.004	0.000	0.394
Education_LevelHigh School Diploma	0.361	0.052	6.890	0.000	1.434
Education_LevelLess than HS diploma	0.199	0.054	3.672	0.000	1.220

Interpretation of the coefficients

- **Age at Release: Age at release 43-47**: An odd ratio 0.652 indicates that individuals aged 43-47 are approximately 0.65 times of the odds of recidivism compared to individuals aged 18-22 (the reference category). In other words, older individuals is associated with a lower likelihood of recidivism as the odd ratio less than 1.
- **Gang Affiliated: Yes**: An odd ratio 3.735 indicates that individuals who are gang affiliated are approximately 3.735 times of the odds of recidivism compared to individuals who are unknown. In other words, gang affiliation is associated with a higher likelihood of recidivism as the odd ratio greater than 1.
- **Percent Days Employed**: An odd ratio 0.253 indicates that individuals with a 1% increase in the percentage of days employed are approximately 0.253 times of the odds of recidivism compared to individuals with a lower percentage of days employed. In other words, higher employment is associated with a lower likelihood of recidivism as the odd ratio less than 1.
- **Education Level: less than high school diploma**: An odd ratio 1.220 indicates that individuals with less than a high school diploma are approximately 1.220 times of the odds of recidivism compared

to individuals with a bachelor degree (the reference category). In other words, lower education level is associated with a higher likelihood of recidivism as the odd ratio greater than 1.

Cuteoff Exploration

Threshold 0.25 and model evaluation

```
test_probs1 <- predict(model, newdata = test, type = "response")
threshold1 <- 0.25
test_preds1 <- ifelse(test_probs1 > threshold1, 1, 0)
confusionMatrix(as.factor(test_preds1), as.factor(test$recidivism_yes), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  156   29
##          1 2320 3337
##
##                Accuracy : 0.5979
##                  95% CI : (0.5852, 0.6105)
##     No Information Rate : 0.5762
##     P-Value [Acc > NIR] : 0.0003946
##
##                   Kappa : 0.062
##
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
##             Sensitivity : 0.9914
##             Specificity : 0.0630
##          Pos Pred Value : 0.5899
##          Neg Pred Value : 0.8432
##              Prevalence : 0.5762
##          Detection Rate : 0.5712
##    Detection Prevalence : 0.9683
##       Balanced Accuracy : 0.5272
##
##        'Positive' Class : 1
##
```

Threshold 0.5 and model evaluation

```
test_probs2 <- predict(model, newdata = test, type = "response")
threshold2 <- 0.5
test_preds2 <- ifelse(test_probs2 > threshold2, 1, 0)
confusionMatrix(as.factor(test_preds2), as.factor(test$recidivism_yes), positive = "1")
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction    0    1
##          0 1337  809
##          1 1139 2557
##
##               Accuracy : 0.6666
##                 95% CI : (0.6543, 0.6786)
##    No Information Rate : 0.5762
##    P-Value [Acc > NIR] : < 0.00000000000000022
##
##                  Kappa : 0.305
##
##  Mcnemar's Test P-Value : 0.00000000000009041
##
##            Sensitivity : 0.7597
##            Specificity : 0.5400
##         Pos Pred Value : 0.6918
##         Neg Pred Value : 0.6230
##             Prevalence : 0.5762
##         Detection Rate : 0.4377
##   Detection Prevalence : 0.6327
##      Balanced Accuracy : 0.6498
##
##       'Positive' Class : 1
##
```

Threshold 0.75 and model evaluation

```r
test_probs3 <- predict(model, newdata = test, type = "response")
threshold3 <- 0.75
test_preds3 <- ifelse(test_probs3 > threshold3, 1, 0)
confusionMatrix(as.factor(test_preds3), as.factor(test$recidivism_yes), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2240 2439
##          1  236  927
##
##               Accuracy : 0.5421
##                 95% CI : (0.5292, 0.5549)
##    No Information Rate : 0.5762
##    P-Value [Acc > NIR] : 1
##
##                  Kappa : 0.1611
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##            Sensitivity : 0.2754
##            Specificity : 0.9047
```

```
##          Pos Pred Value : 0.7971
##          Neg Pred Value : 0.4787
##              Prevalence : 0.5762
##          Detection Rate : 0.1587
##    Detection Prevalence : 0.1991
##       Balanced Accuracy : 0.5900
##
##        'Positive' Class : 1
##
```

Threshold Comparison and AUC

Table 2: Comparison of Model Performance at Different Thresholds

	Threshold	Accuracy	Sensitivity	Specificity
Threshold 1	0.25	0.598	0.991	0.063
Threshold 2	0.50	0.667	0.760	0.540
Threshold 3	0.75	0.542	0.275	0.905

```
roc_obj <- roc(test$recidivism_yes, test_probs1)
plot(roc_obj, col = "blue", main = "ROC Curve with AUC", print.auc = TRUE, print.auc.x = 0.5, print.auc
```


ROC Curve with AUC

AUC: 0.713