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

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

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

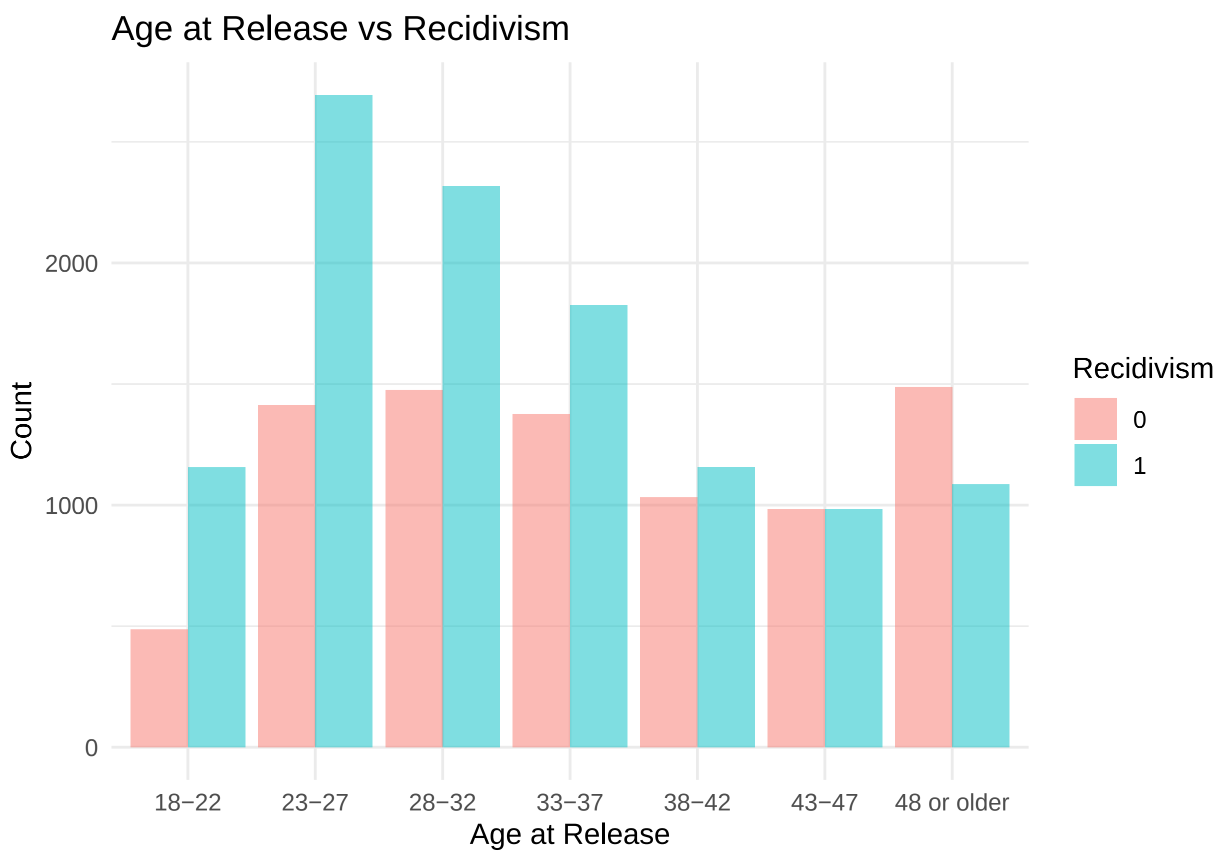
Training and Testing Partition

```r
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

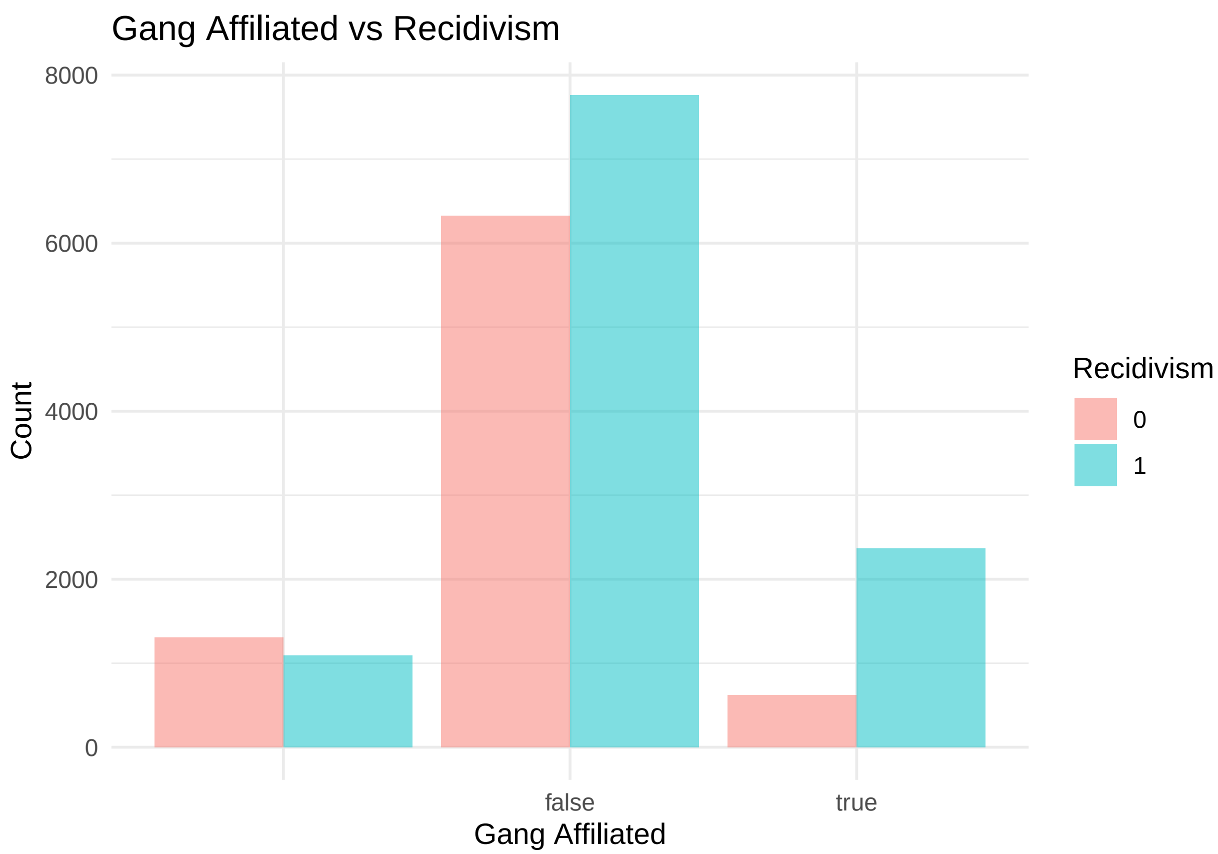
- **Age at Release**: Age at release is a significant predictor of recidivism. Younger individuals are more likely to re-offend compared to older individuals.

```r
ggplot(recidivism_ga, aes(x = Age_at_Release, fill = factor(recidivism_yes))) +
  geom_bar(stat = "count", position = "dodge", alpha = 0.5) +
  labs(title = "Age at Release vs Recidivism",
       x = "Age at Release",
       y = "Count",
       fill = "Recidivism") +
  theme_minimal()
```



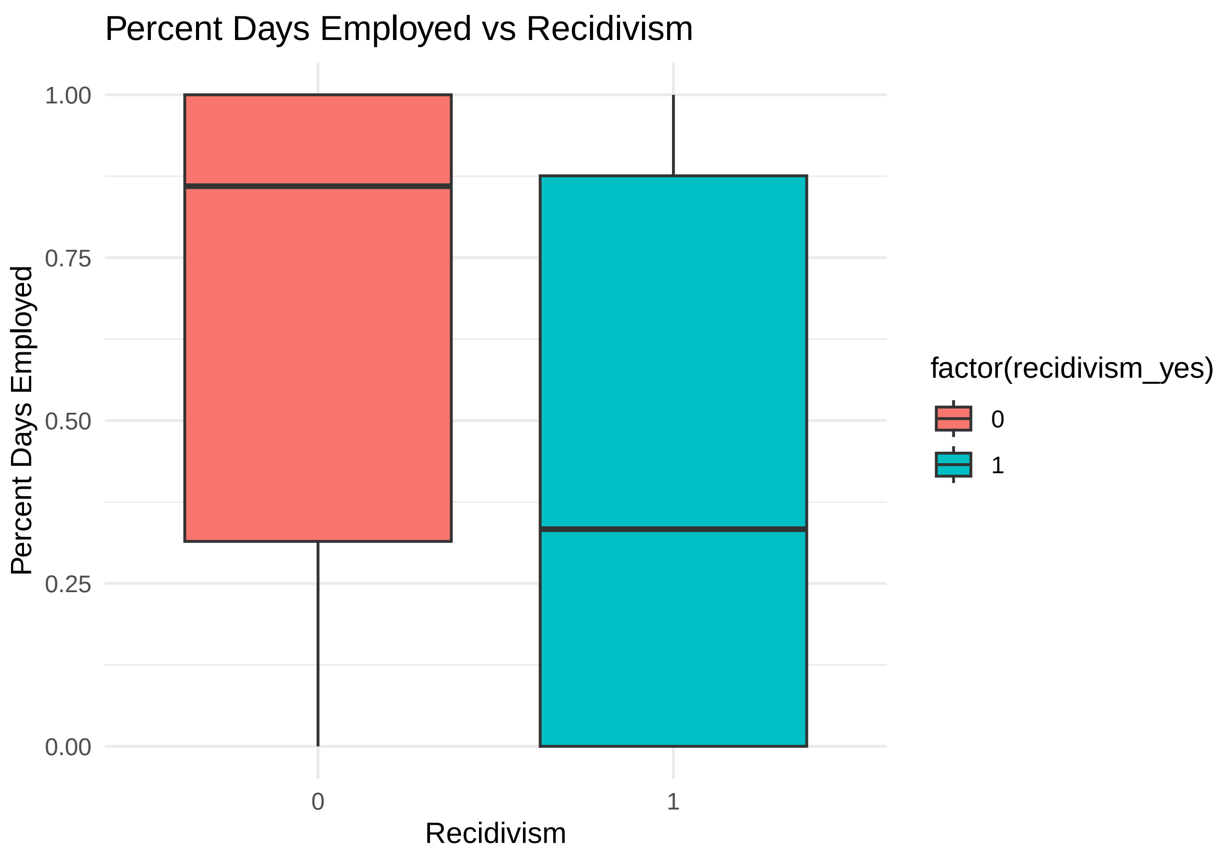
- **Gang Affiliated**: Gang affiliation is a significant predictor of recidivism. Individuals who are affiliated with gangs are more likely to re-offend compared to those who are not.

```
ggplot(recidivism_ga, aes(x = Gang_Affiliated, fill = factor(recidivism_yes))) +
  geom_bar(stat = "count", position = "dodge", alpha = 0.5) +
  labs(title = "Gang Affiliated vs Recidivism",
       x = "Gang Affiliated",
       y = "Count",
       fill = "Recidivism") +
  theme_minimal()
```



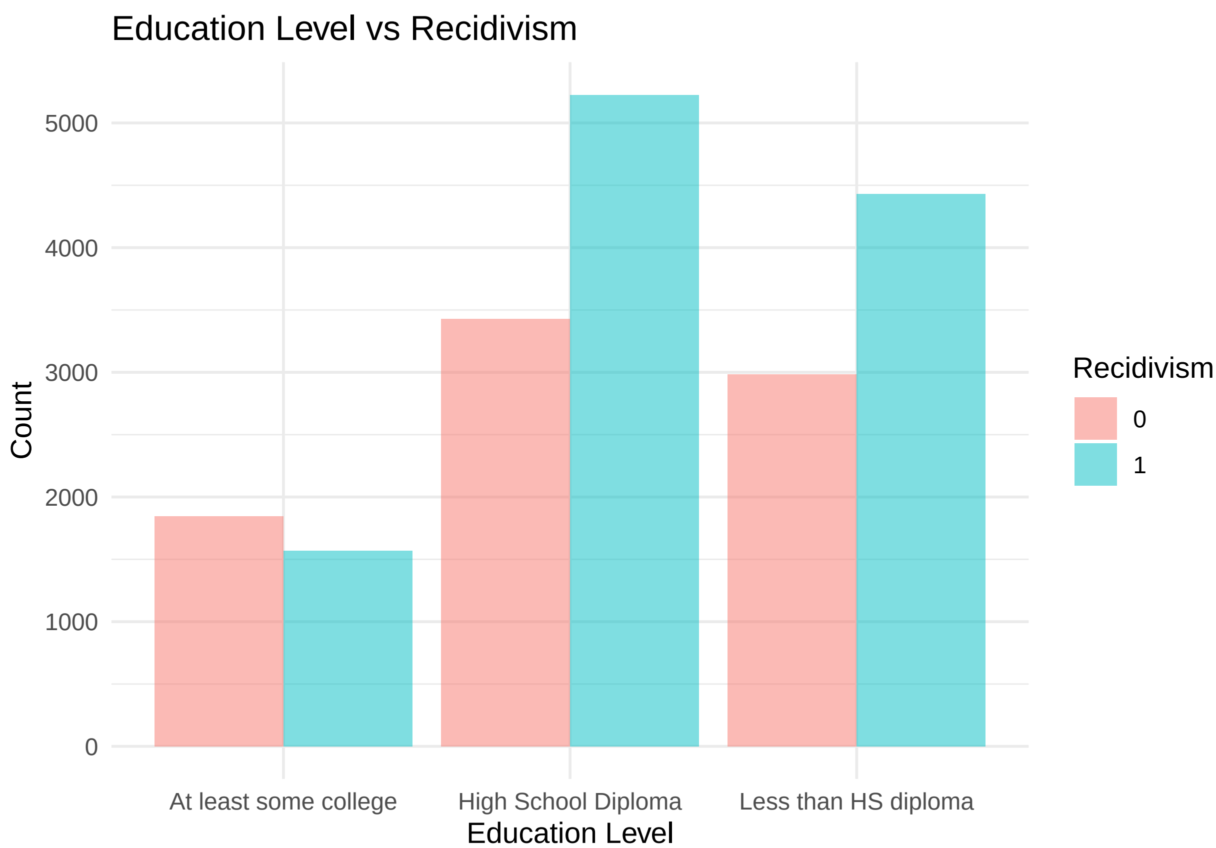
- **Percent Days Employed**: Percent days employed is a significant predictor of recidivism. Individuals who are employed for a higher percentage of days are less likely to re-offend compared to those who are unemployed.

```
ggplot(recidivism_ga, aes(x = factor(recidivism_yes), y = Percent_Days_Employed, fill = factor(recidivis
  geom_boxplot() +
  labs(title = "Percent Days Employed vs Recidivism",
       x = "Recidivism",
       y = "Percent Days Employed") +
  theme_minimal()
```



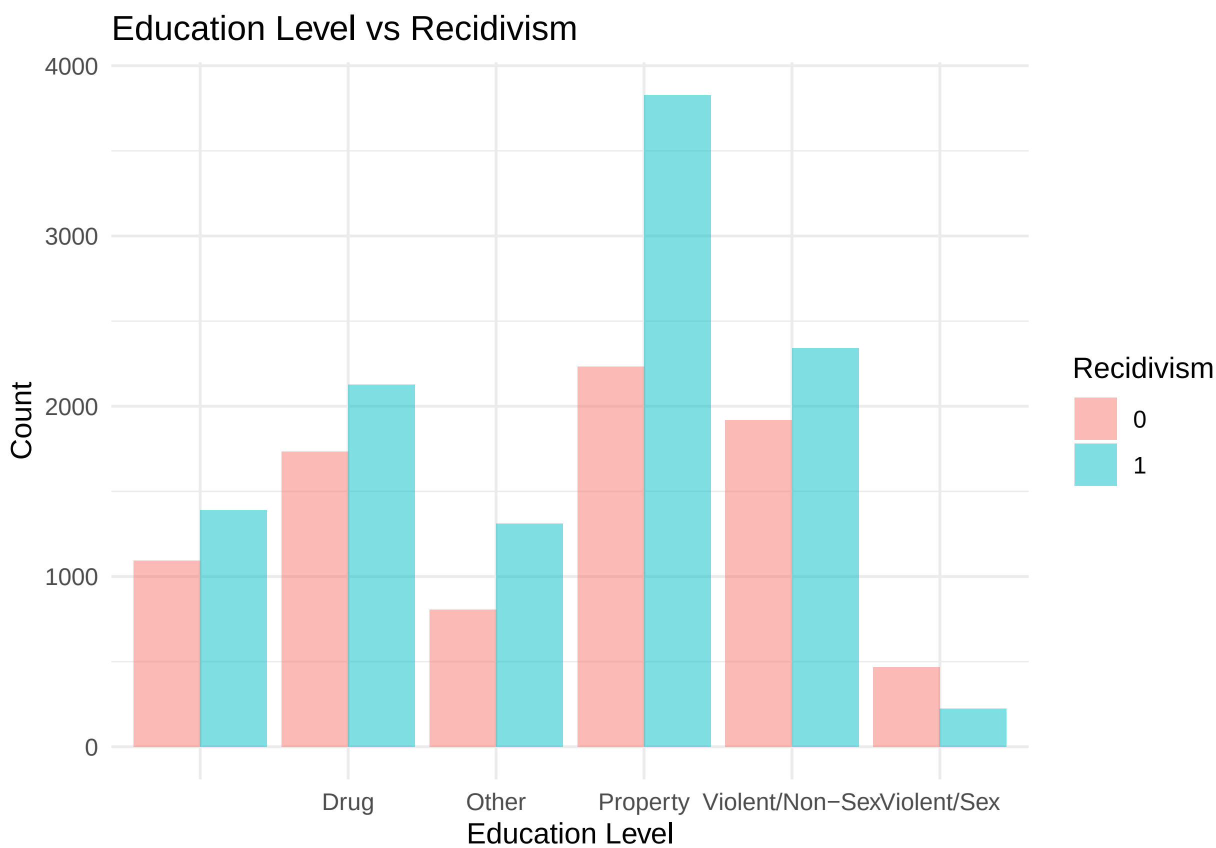
Education Level: Education level is a significant predictor of recidivism. Individuals with higher education levels are less likely to re-offend compared to those with lower education levels.

```
ggplot(recidivism_ga, aes(x = Education_Level, fill = factor(recidivism_yes))) +
  geom_bar(stat = "count", position = "dodge", alpha = 0.5) +
  labs(title = "Education Level vs Recidivism",
       x = "Education Level",
       y = "Count",
       fill = "Recidivism") +
  theme_minimal()
```



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.

```
ggplot(recidivism_ga, aes(x = Prison_Offense, fill = factor(recidivism_yes))) +
  geom_bar(stat = "count", position = "dodge", alpha = 0.5) +
  labs(title = "Education Level vs Recidivism",
       x = "Education Level",
       y = "Count",
       fill = "Recidivism") +
  theme_minimal()
```



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")
summary(model)
```

```
##
## Call:
## glm(formula = recidivism_yes ~ Age_at_Release + Gang_Affiliated +
##     Percent_Days_Employed + Prison_Offense + Education_Level,
##     family = "binomial", data = train)
##
## Coefficients:
##                                        Estimate Std. Error z value
## (Intercept)                             0.48067    0.10753   4.470
## Age_at_Release23-27                    -0.03234    0.08092  -0.400
## Age_at_Release28-32                    -0.10013    0.08197  -1.222
## Age_at_Release33-37                    -0.16872    0.08412  -2.006
## Age_at_Release38-42                    -0.29611    0.08890  -3.331
## Age_at_Release43-47                    -0.42720    0.09149  -4.669
## Age_at_Release48 or older              -0.78251    0.08728  -8.966
```

```
## Gang_Affiliatedfalse                0.46365    0.05606   8.271
## Gang_Affiliatedtrue                 1.31768    0.07789  16.918
## Percent_Days_Employed              -1.37484    0.04648 -29.579
## Prison_OffenseDrug                 -0.01852    0.06632  -0.279
## Prison_OffenseOther                 0.19645    0.07726   2.543
## Prison_OffenseProperty              0.33881    0.06243   5.427
## Prison_OffenseViolent/Non-Sex      -0.11061    0.06567  -1.684
## Prison_OffenseViolent/Sex          -0.93217    0.11647  -8.004
## Education_LevelHigh School Diploma  0.36051    0.05233   6.890
## Education_LevelLess than HS diploma  0.19908    0.05421   3.672
##                                               Pr(>|z|)
## (Intercept)                         0.00000781695385641 ***
## Age_at_Release23-27                           0.689394
## Age_at_Release28-32                           0.221864
## Age_at_Release33-37                           0.044879 *
## Age_at_Release38-42                           0.000865 ***
## Age_at_Release43-47                 0.00000302094613714 ***
## Age_at_Release48 or older          < 0.0000000000000002 ***
## Gang_Affiliatedfalse               < 0.0000000000000002 ***
## Gang_Affiliatedtrue                < 0.0000000000000002 ***
## Percent_Days_Employed              < 0.0000000000000002 ***
## Prison_OffenseDrug                            0.780078
## Prison_OffenseOther                           0.010999 *
## Prison_OffenseProperty              0.00000005735005517 ***
## Prison_OffenseViolent/Non-Sex                 0.092116 .
## Prison_OffenseViolent/Sex           0.00000000000000121 ***
## Education_LevelHigh School Diploma  0.00000000000559572 ***
## Education_LevelLess than HS diploma           0.000241 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 18586  on 13635  degrees of freedom
## Residual deviance: 16578  on 13619  degrees of freedom
## AIC: 16612
##
## Number of Fisher Scoring iterations: 4
```