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

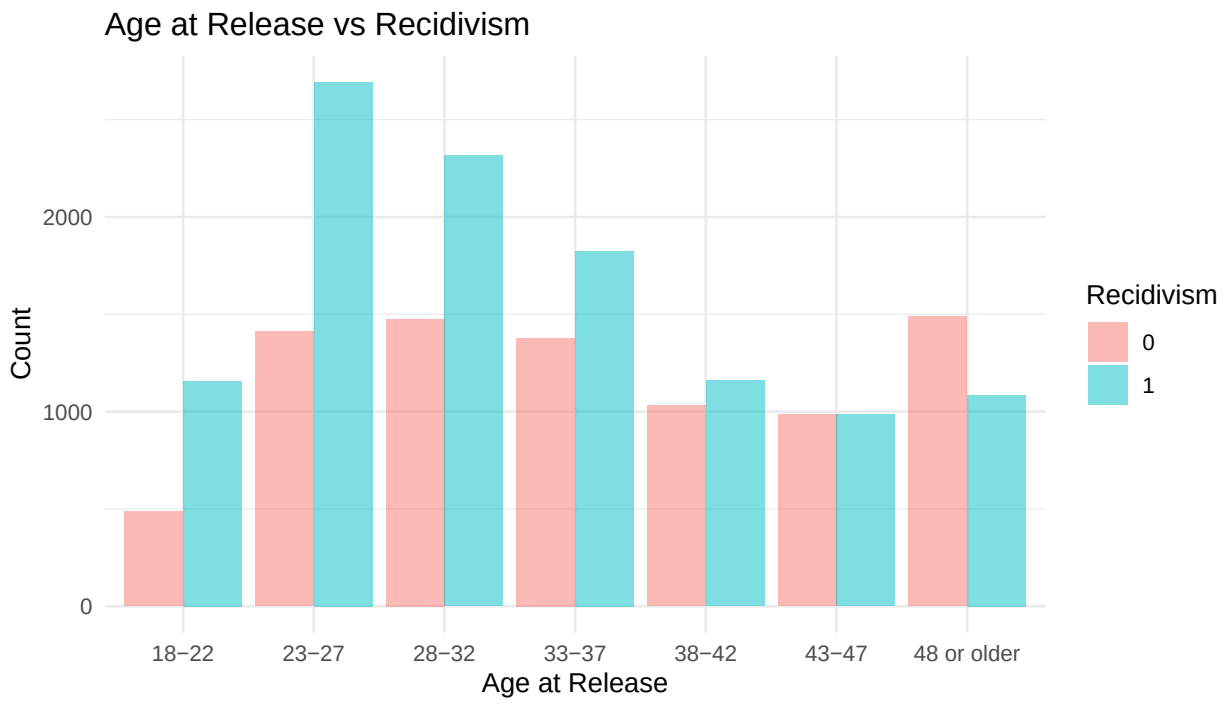
```r
recidivism_ga <- data %>%
  mutate(recidivism_yes = if_else(Recidivism_Within_3years == "true", 1, 0))

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

Training and Testing Partition

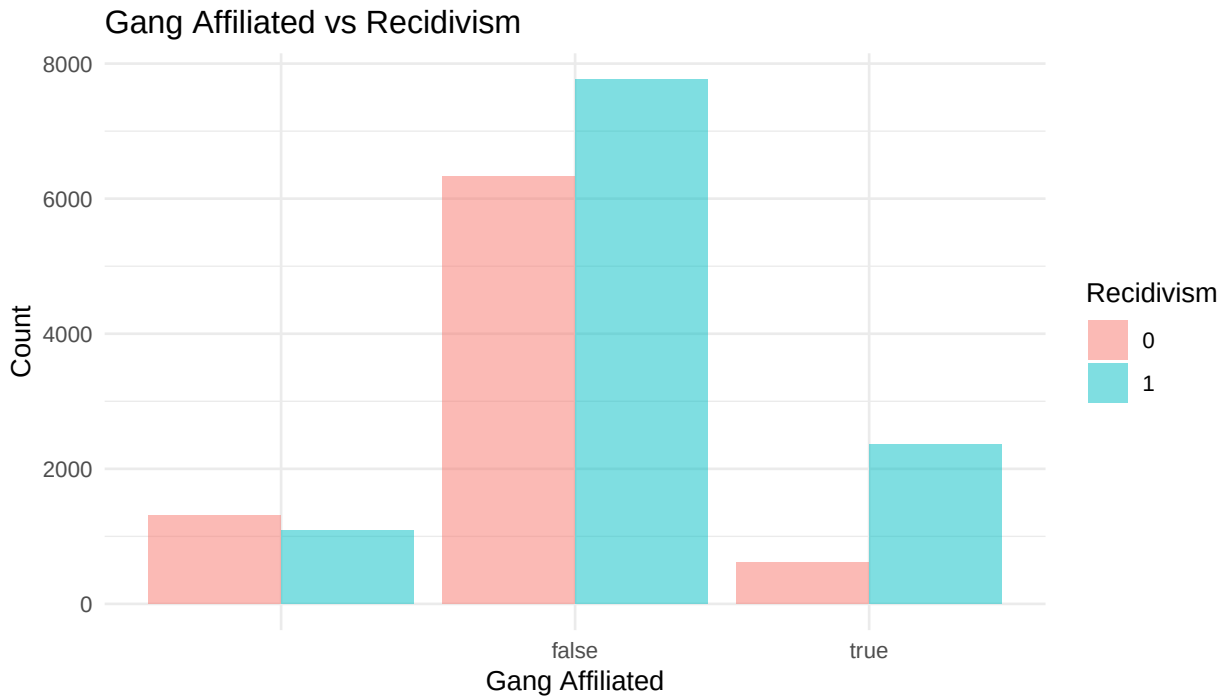
```r
set.seed(1234)
trainIndex <- createDataPartition(as.factor(recidivism_ga$recidivism_yes), p = 0.7, list = FALSE)
train <- recidivism_ga[trainIndex, ]
test <- recidivism_ga[-trainIndex, ]
```

Key predictors

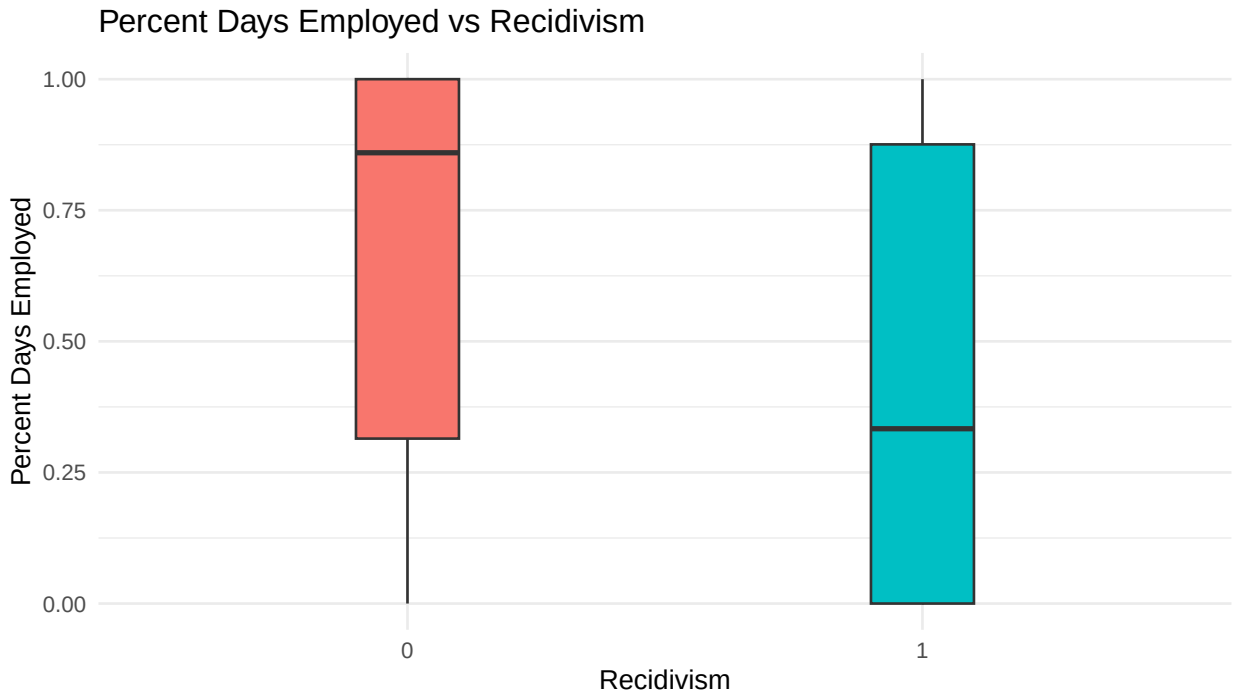
- **Age at Release**: Age at release is a significant predictor of recidivism. Younger individuals are more likely to re-offend compared to older individuals.



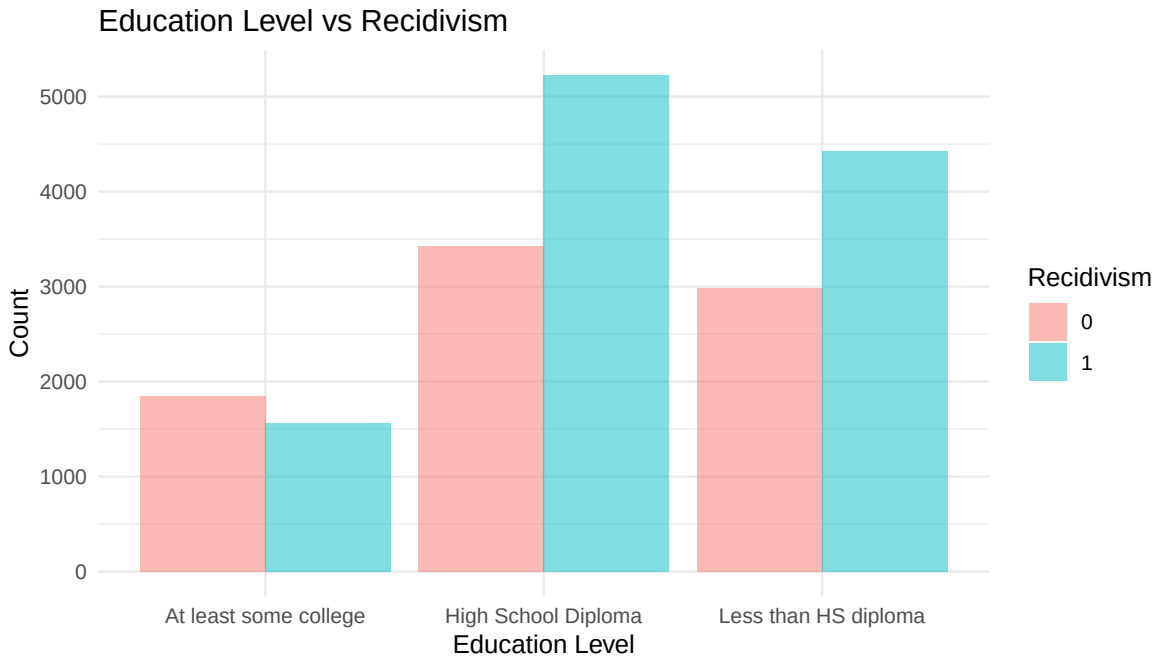
- **Gang Affiliated**: Gang affiliation is a significant predictor of recidivism. Individuals who are affiliated with gangs are more likely to re-offend compared to those who are not.



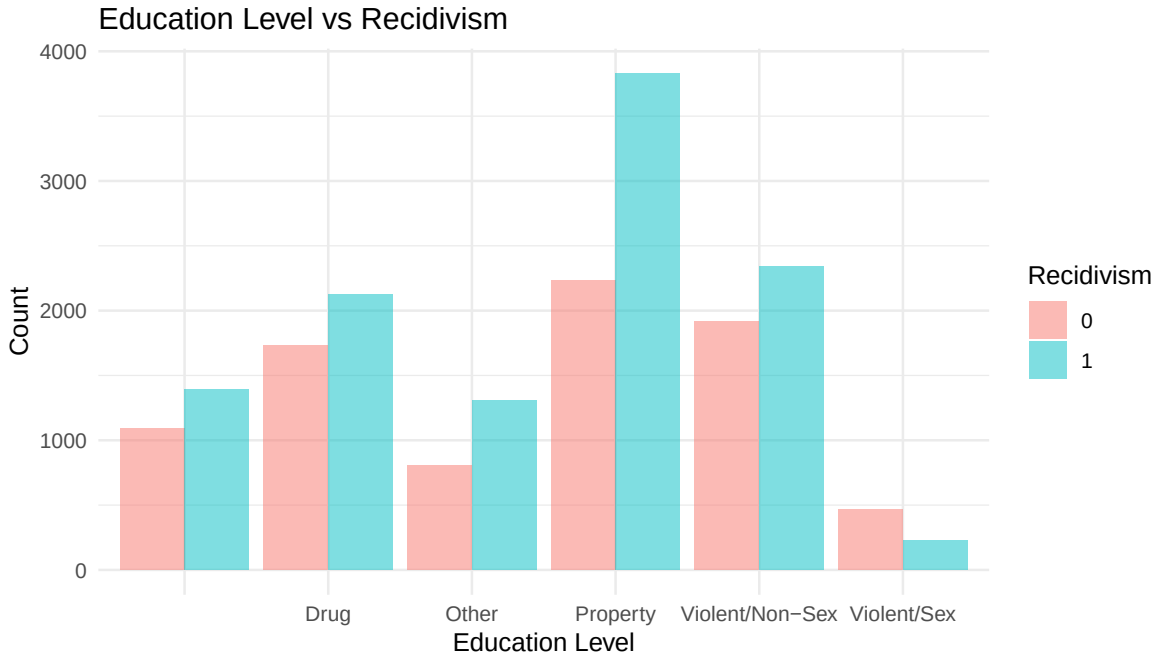
- **Percent Days Employed**: Percent days employed is a significant predictor of recidivism. Individuals who are employed for a higher percentage of days are less likely to re-offend compared to those who are unemployed.



- **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.



- **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.



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

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

Table 1: Logistic Regression Coefficient Summary

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