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

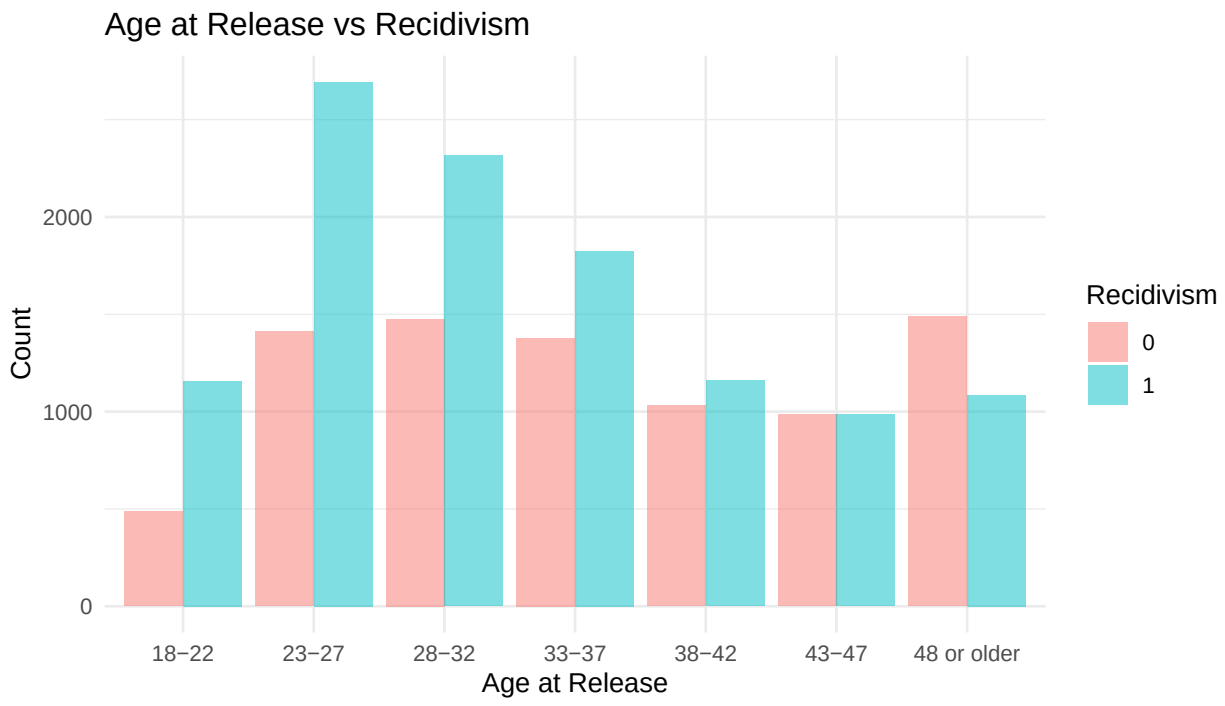
```r
recidivism_ga <- data %>%
  mutate(recidivism_yes = if_else(Recidivism_Within_3years == "true", 1, 0))

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

Training and Testing Partition

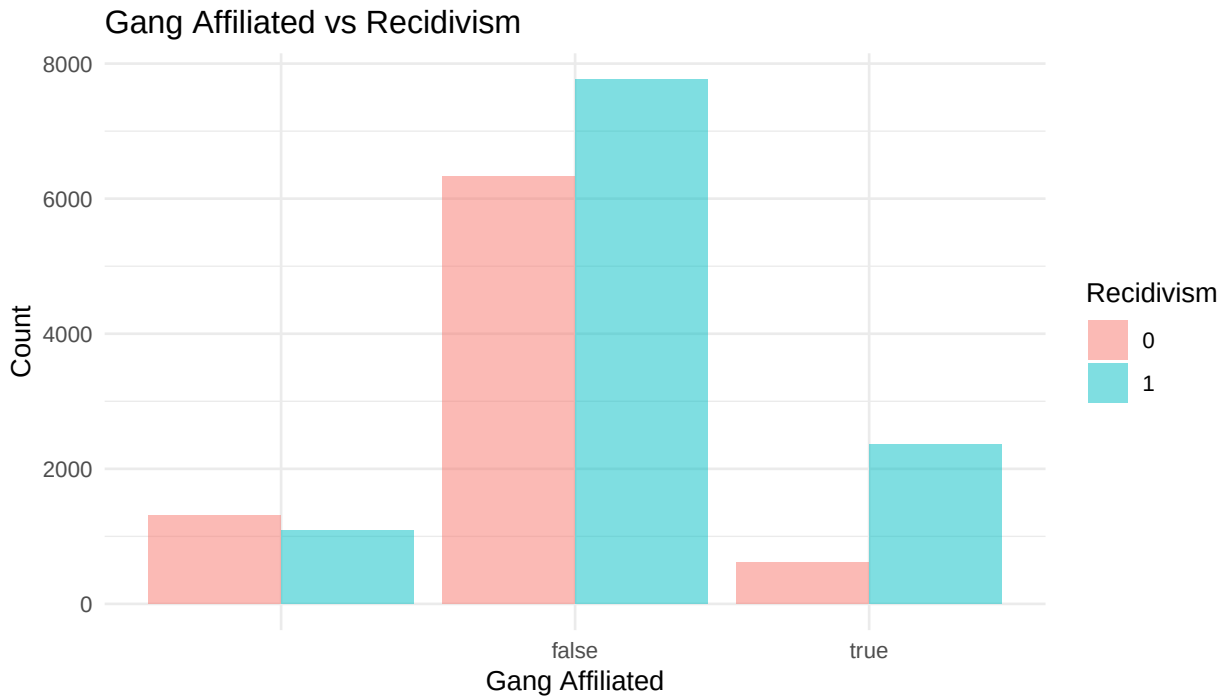
```r
set.seed(1234)
trainIndex <- createDataPartition(as.factor(recidivism_ga$recidivism_yes), p = 0.7, list = FALSE)
train <- recidivism_ga[trainIndex, ]
test <- recidivism_ga[-trainIndex, ]
```

Key predictors

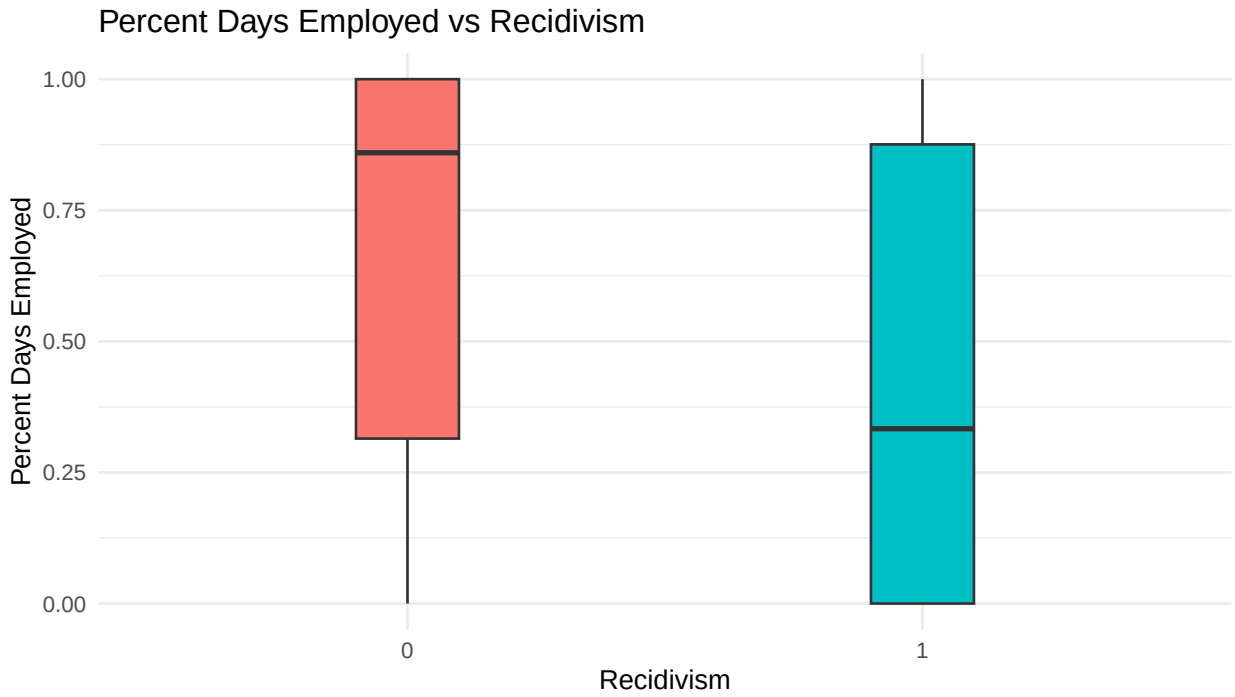
- **Age at Release**: Age at release is a significant predictor of recidivism. Younger individuals are more likely to re-offend compared to older individuals.



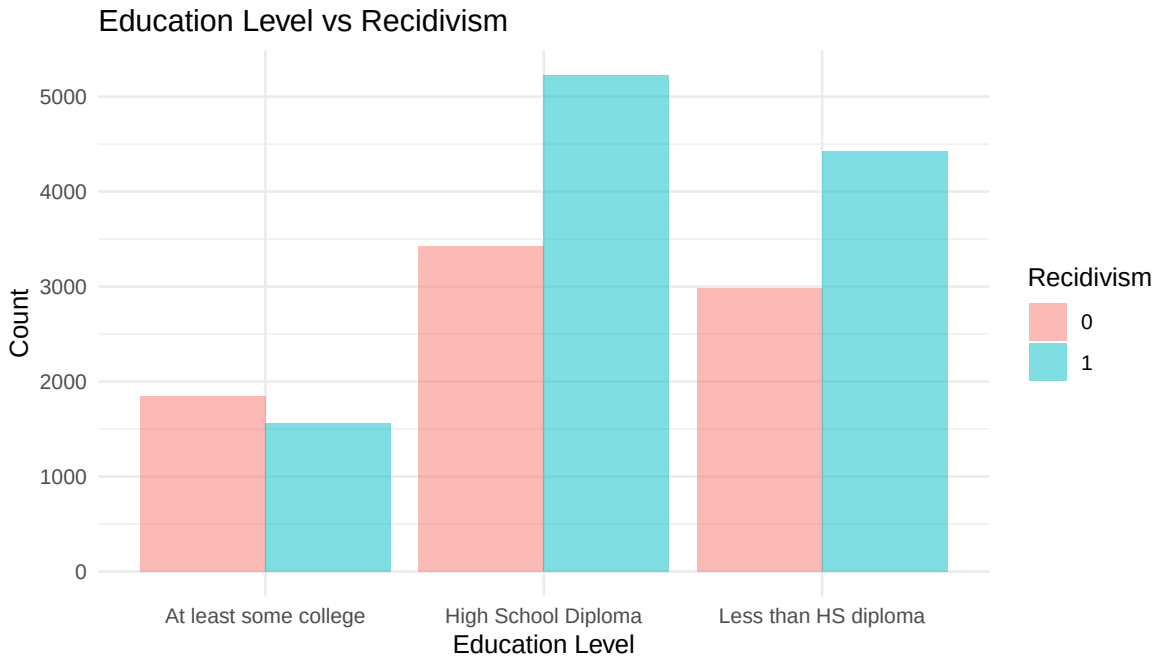
- **Gang Affiliated**: Gang affiliation is a significant predictor of recidivism. Individuals who are affiliated with gangs are more likely to re-offend compared to those who are not.



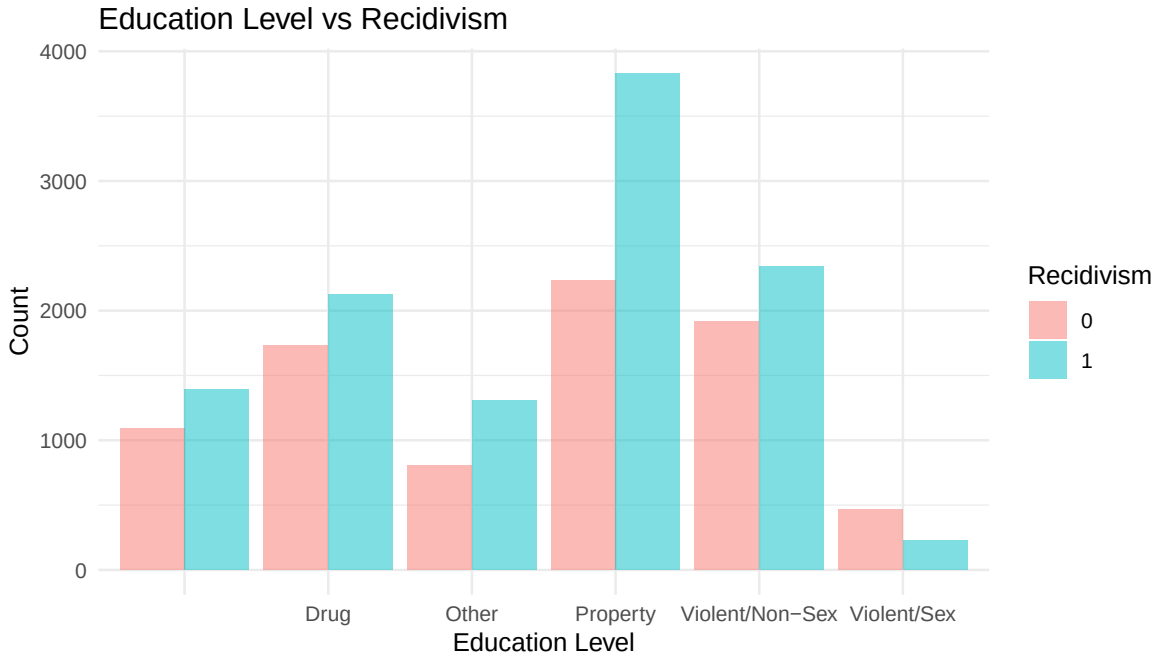
- **Percent Days Employed**: Percent days employed is a significant predictor of recidivism. Individuals who are employed for a higher percentage of days are less likely to re-offend compared to those who are unemployed.



- **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

# 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

In this section, I test different thresholds to see how the model performs. I use 0.25, 0.5, and 0.75 as the thresholds to see how the model performs. Different thresholds lead to different balance between sensitivity and specificity. I use the confusion matrix to evaluate the model performance at different thresholds.

Threshold 0.25 and model evaluation

Confusion matrix for the threshold 0.25 is shown below. The model has a sensitivity of 0.99, specificity of 0.06, and a 0.598 overall precision. This model presents a high sensitivity, meaning it is very good at identifying individuals who are likely to re-offend (true positive). However, the specificity is low, indicating that many individuals who are not likely to re-offend are incorrectly classified as recidivists (true negative). This model is useful for identifying high-risk individuals but may lead to unnecessary detentions of low-risk individuals.

```
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

Confusion matrix for the threshold 0.5 is shown below. In contrast with the first one, the 0.5 threshold provide a balance between the sensitivity and specificity. Thus, the model also has a better precision in overall prediction. The model has a sensitivity of 0.75, specificity of 0.5400 , and a 0.666 overall precision.

```
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

Confusion matrix for the threshold 0.75 is shown below. The model has a sensitivity of 0.28, specificity of 0.90, and a 0.54 overall precision. This model presents a high specificity, meaning it is very good at identifying individuals who are not likely to re-offend (true negative). However, the sensitivity is low, indicating that many individuals who are likely to re-offend are incorrectly classified as non-recidivists (true positive). This model is useful for identifying low-risk individuals but may lead to release of high-risk individuals who likely to re-offend.

```
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

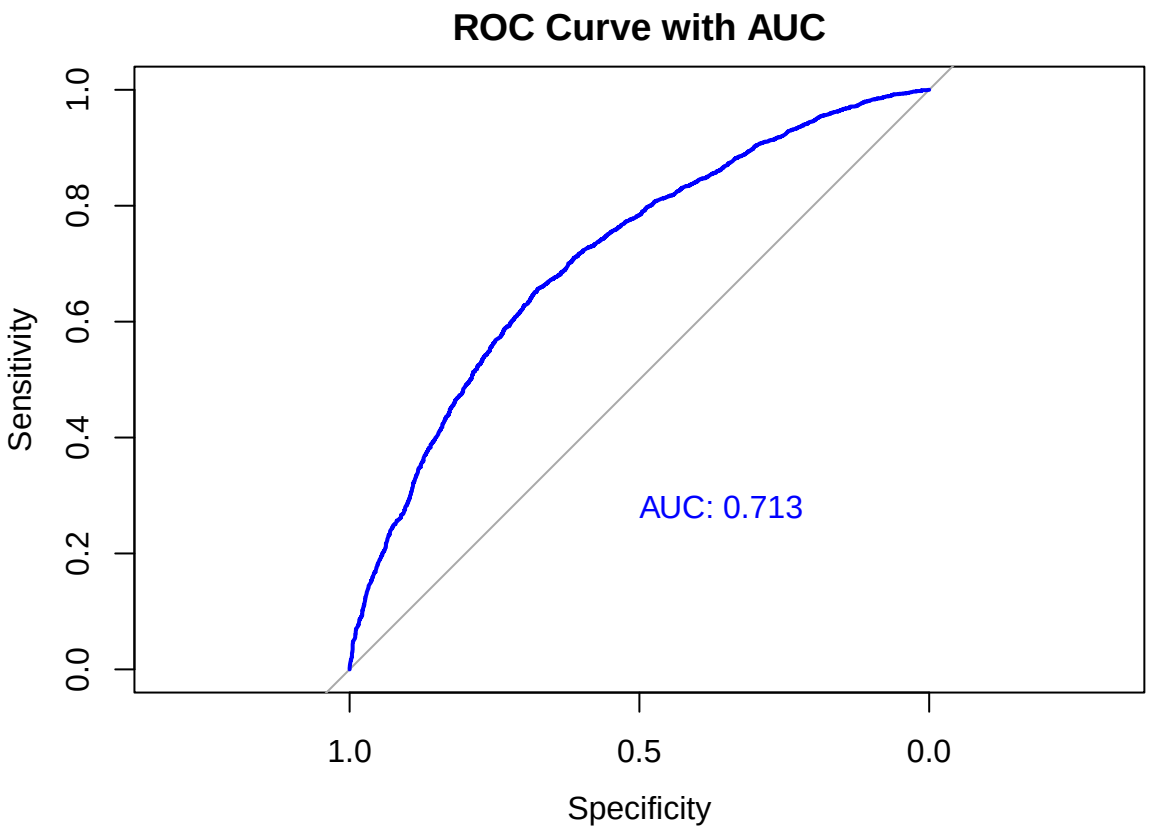
Overall, adjust the threshold adjusts the balance between sensitivity and specificity. In general, when set the threshold bigger, the sensitivity decreases and the specificity increases. The threshold of 0.25 provides a high sensitivity but low specificity. The threshold of 0.5 provides a balance between sensitivity and specificity. The threshold of 0.75 provides a high specificity but low sensitivity. In this specific case, I would prefer to the first or the second model threshold, as the first priority is to prevent the criminal to re-offend and efficient using our limited detention spaces.

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

The overall AUC of the model is 0.75, which indicates that the model has a good ability to distinguish between recidivists and non-recidivists. It also shows the different cut-off values and their corresponding sensitivity and specificity.

```
roc_obj <- roc(test$recidivism_yes, test_probs2)
plot(roc_obj, col = "blue", main = "ROC Curve with AUC", print.auc = TRUE, print.auc.x = 0.5, print.auc
```



Equity Audit by Race and Age Group

Equity Audit by Race

Our model didn't include the race as predictors to avoid bias. However, the results of the matrix still indicate that the model perform dramatically differently regarding people from different race group.

As shown in the table below, the model perform similarly in overall precision (around 70%). However, the model has a relative high sensitivity and low specificity in predicting the black population than the white population. The model has a sensitivity of 0.802 and specificity of 0.486 which indicates the black population group are more likely to be misclassified as recidivists (higher true positive). Black population are more likely to detained rather than release. In contrast, the model has a sensitivity of 0.70 and specificity of 0.61 in predicting the white population group. This indicates that the chances for white population to be misclassfied as recidivists is lower than black population. Although we didn;t include any race related predictors in the model, the model still present a bias against the black population group.

```
group_metric_summary <- test %>%
  mutate(pred = test_preds2, prob = test_probs2) %>%
  group_by(Race) %>%
  summarise(
    Sensitivity = round(sum(pred == 1 & recidivism_yes == 1) / sum(recidivism_yes == 1), 3),
    Specificity = round(sum(pred == 0 & recidivism_yes == 0) / sum(recidivism_yes == 0), 3),
    Precision   = round(sum(pred == 1 & recidivism_yes == 1) / sum(pred == 1), 3),
    Number = n()
```

```
  )
kable(group_metric_summary, caption = "Equity Audit by Race")
```

Table 3: Equity Audit by Race

Race	Sensitivity	Specificity	Precision	Number
BLACK	0.802	0.486	0.690	3364
WHITE	0.700	0.609	0.695	2478

Equity Audit by Age Group

The model include the age group as a predictors as younger population are more likely to re-offend. For the precision,the model perform well when the people age is younger than older, as indicated by high precision rate (0.71 vs 0.55). In addition, younger population group has a relative higher sensitivity and lower specificity than older population group. The model has a sensitivity of 0.903 and specificity of 0.176 in predicting the younger population group (18-22). This indicates that the younger population group are more likely to be misclassified as recidivists (higher true positive). In other words, younger population group are more likely to be detained rather than release. In contrast, the model has a sensitivity of 0.503 and specificity of 0.709 in predicting the older population group (48 or older). This indicates that the chances for older population to be misclassfied as non-recidivists is higher. In other words, older population group are more likely to be release rather than detained. This also indicates the model is potentially biased against the younger population group and have more room to improve.

```
group_metric_summary2 <- test %>%
  mutate(pred = test_preds2, prob = test_probs2) %>%
  group_by(Age_at_Release) %>%
  summarise(
    Sensitivity = round(sum(pred == 1 & recidivism_yes == 1) / sum(recidivism_yes == 1), 3),
    Specificity = round(sum(pred == 0 & recidivism_yes == 0) / sum(recidivism_yes == 0), 3),
    Precision   = round(sum(pred == 1 & recidivism_yes == 1) / sum(pred == 1), 3),
    Number = n()
  )

kable(group_metric_summary2 %>% rename(Age=Age_at_Release), caption = "Equity Audit by Age Group")
```

Table 4: Equity Audit by Age Group

Age	Sensitivity	Specificity	Precision	Number
18-22	0.903	0.176	0.716	488
23-27	0.869	0.399	0.736	1185
28-32	0.824	0.453	0.710	1166
33-37	0.738	0.501	0.652	1009
38-42	0.643	0.669	0.692	606
43-47	0.606	0.721	0.687	620
48 or older	0.503	0.709	0.555	768

Cost-Benefit Analysis

Step 1; Estimate the cost

Scnerio 1: Cost of recidivism (released and then reoffernds)

A false negative occurs when an individual is predicted to be low risks and released. However, then commits another offense after release. There are two major costs associated with this scenario: - **Direct Criminal Justice Costs**: This includes the costs associated with criminal justice system, including police investigation, prosecutions, and trail-related expenses. - **Victim and Social Costs**: this includes the costs associated with the crime itself, including property damage, medical expenses, lost productivity, and even people lose their life. In addition, there are additional social influences , like people lose trusts to the justice system, a places where the crime happen faces economic decline, and the community is less safe. - **Total Cost**: The total cost of recidivism is the sum of the direct criminal justice costs and the victim and social costs is hard to estimate. The average cost of recidivism is estimated to be around $100,000 per individual, but also varies between crime types.

Scenario 2: Cost of false positive (detained someone in jail unnecessarily)

A false positive occurs when an individual is predicted to be high risks and detained. However, they were misclassfied as high risks and not likely to re-offense. Only one costs associated with this scenario: - **Cost of Detention**: This includes the costs associated with detaining an individual in jail, including housing, food, medical care, and other expenses. The average cost of detention is estimated to be around $30,000 per year per individual. Depends on prosecution timeline, the average time in jail for suspect waiting for trail is also varies.

Step 2: Estimate the benefits

Scenario 1: Benefits of preventing recidivism

The benefits of preventing recidivism to make sure the individuals who are likely to re-offend are not released include: - Avoid extra cost of re-offense - Community become safer

Scenario 2: Benefits of releasing low-risk individuals

The benefits of releasing low-risk individuals include: - Government save costs on extra detention costs - People are released can create economic benefits to the society (go to work and make money)

Step 3: Balancing costs and cutoff

The best way to balance the costs is to minimize the total costs. Basically, the total costs is the sum of the costs associated with false negatives and false positives. The total cost can be calculated as follows:

$$\text{Total Cost} = (FNR \cdot \text{Cost}_{\text{Recidivism}}) + (FPR \cdot \text{Cost}_{\text{Detention}})$$

Where: - FNR: False Negative Rate - FPR: False Positive Rate - Cost_Recidivism: Cost of recidivism - Cost_Detention: Cost of detention

Since the costs associated with false negative are roughly 3- 3.5 times higher than the false positive, the cutoff should set to be more conservatively (less than 0.5 and greater 0.25).

Policy recommendation

I recommend the government to set the cutoff threshold to 0.5 for following reason:

- **Overall performance**: The threshold of 0.5 provides a balance between sensitivity and specificity and an overall greater precision compared to threshold 0.25 and 0.75.
- **Economic benefits**: The threshold of 0.5 provides a balance between the costs associated with false negatives and false positives. The government can save costs on extra detention costs and avoid the costs associated with recidivism.
- **Equity between different groups**: The threshold of 0.5 provides a general balance between specificity and sensitivity between racial group. First, the model not include the racial group as a predictors to avoid racial bias.Second, although the black population is more likely to be detained based on the model performance, the threshold of 0.5 provides a generalize balance between detention and release (0.802 vs. 0.486). Since the cost associate the false release would be higher, the difference between sensitivity and specificity is acceptable. In addition, since age group at release are included as a predictors, the younger age group are more likely to be detained based on the model performance. However, younger generation is more likely to re-offend based on the data, the model is acceptable.