Preliminary Analysis

Debang Ou

2024-08-29

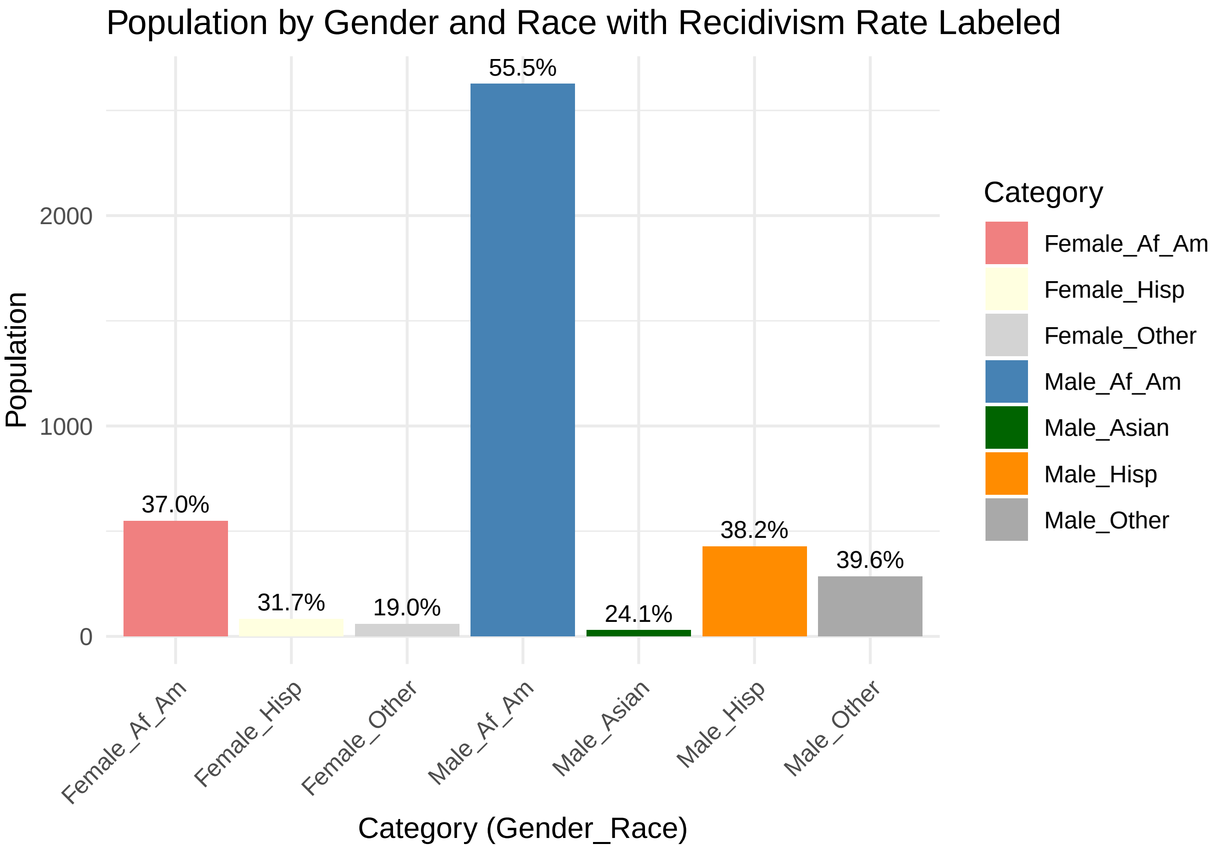
First, we remove rows without proper race indicators.

```r
dataset <- dataset %>%
  filter(rowSums(select(., af_am, asian, hisp, nat_am, other)) == 1)
```

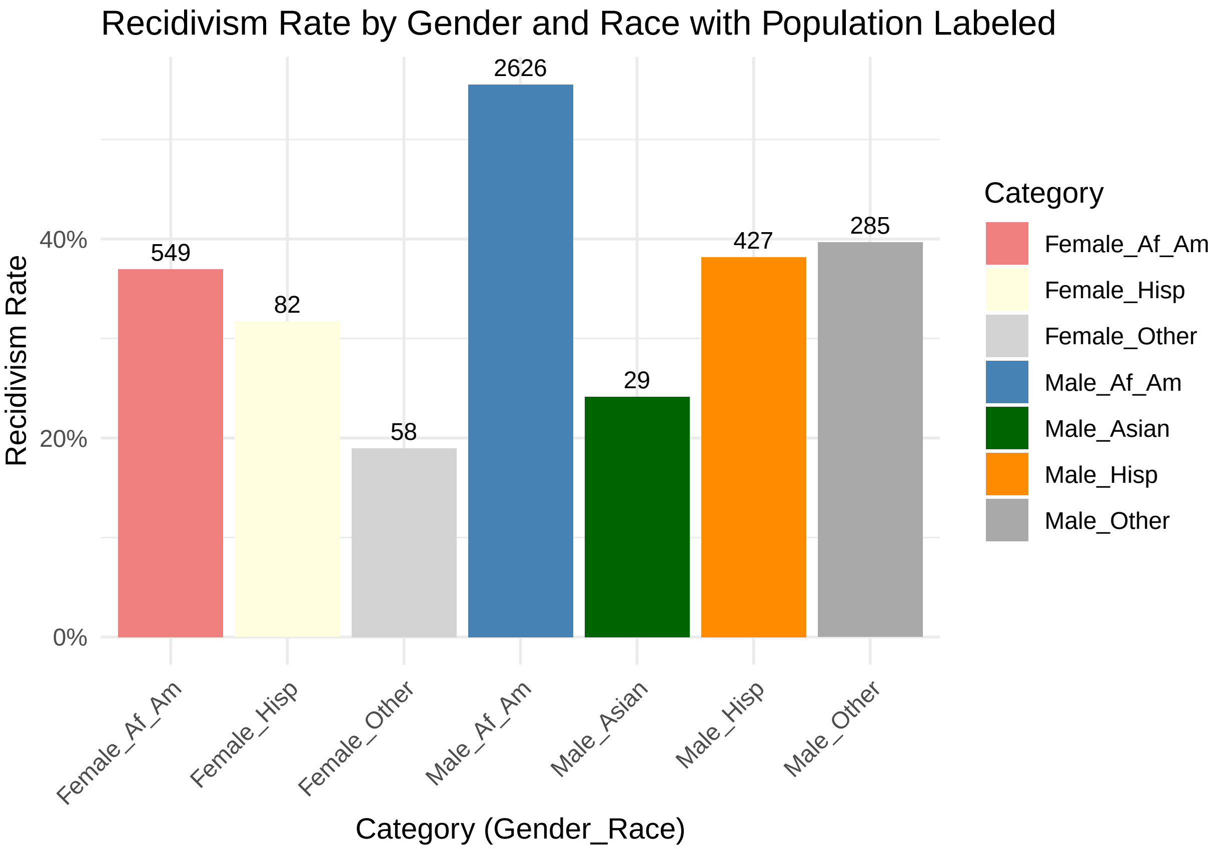
We then visualize the dataset to explore how to start modeling.

Distribution of recidivism and population

```r
# Generate combination categories of gender and race
dataset_combined <- dataset %>%
  mutate(category = factor(case_when(
    female == 1 & af_am == 1 ~ "Female_Af_Am",
    female == 0 & af_am == 1 ~ "Male_Af_Am",
    female == 1 & asian == 1 ~ "Female_Asian",
    female == 0 & asian == 1 ~ "Male_Asian",
    female == 1 & hisp == 1 ~ "Female_Hisp",
    female == 0 & hisp == 1 ~ "Male_Hisp",
    female == 1 & nat_am == 1 ~ "Female_Nat_Am",
    female == 0 & nat_am == 1 ~ "Male_Nat_Am",
    female == 1 & other == 1 ~ "Female_Other",
    female == 0 & other == 1 ~ "Male_Other"
  )))

# Calculate the population for each category
category_recidivism <- dataset_combined %>%
  group_by(category) %>%
  summarise(
    recid_rate = mean(recid),
    population = n()
  )

# Remove categories with population less than 10
#(may be a bad idea?)
category_recidivism <- category_recidivism %>%
  filter(population >= 10)

# Population plot with recidivism rate labeled
ggplot(category_recidivism, aes(x = category, y = population, fill = category)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = scales::percent(recid_rate, accuracy = 0.1)),
  vjust = -0.5, size = 3) +labs(x = "Category (Gender_Race)",
  y = "Population", fill = "Category") +
  ggtitle("Population by Gender and Race with Recidivism Rate Labeled") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(values = c(
    "Female_Af_Am" = "lightcoral", "Male_Af_Am" = "steelblue",
    "Female_Asian" = "lightgreen", "Male_Asian" = "darkgreen",
    "Female_Hisp" = "lightyellow", "Male_Hisp" = "darkorange",
    "Female_Nat_Am" = "pink", "Male_Nat_Am" = "darkred",
    "Female_Other" = "lightgray", "Male_Other" = "darkgray"
  ))
```

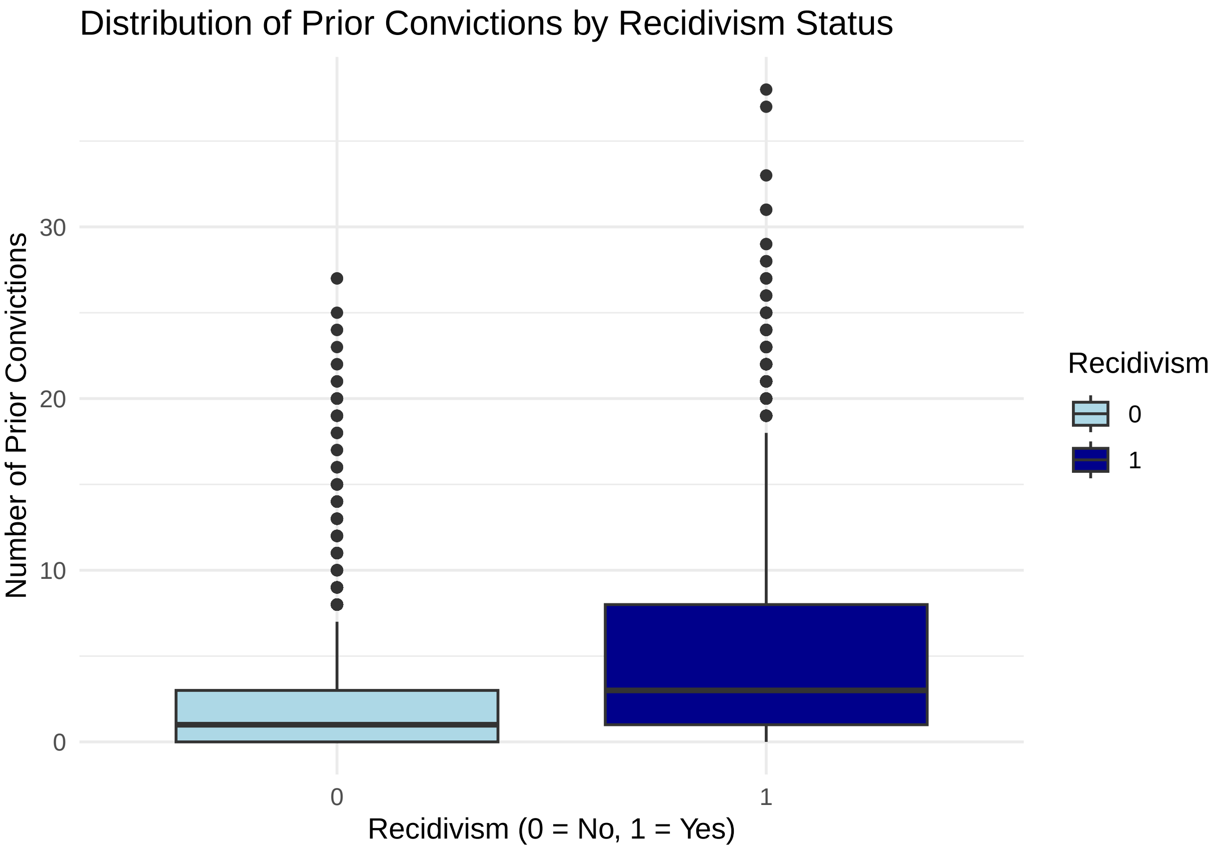


```
# Recidivism rate plot with population labeled
ggplot(category_recidivism, aes(x = category, y = recid_rate, fill = category)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = population), vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Category (Gender_Race)", y = "Recidivism Rate", fill = "Category") +
  ggtitle("Recidivism Rate by Gender and Race with Population Labeled") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(values = c(
    "Female_Af_Am" = "lightcoral", "Male_Af_Am" = "steelblue",
    "Female_Asian" = "lightgreen", "Male_Asian" = "darkgreen",
    "Female_Hisp" = "lightyellow", "Male_Hisp" = "darkorange",
    "Female_Nat_Am" = "pink", "Male_Nat_Am" = "darkred",
    "Female_Other" = "lightgray", "Male_Other" = "darkgray"
  ))
```

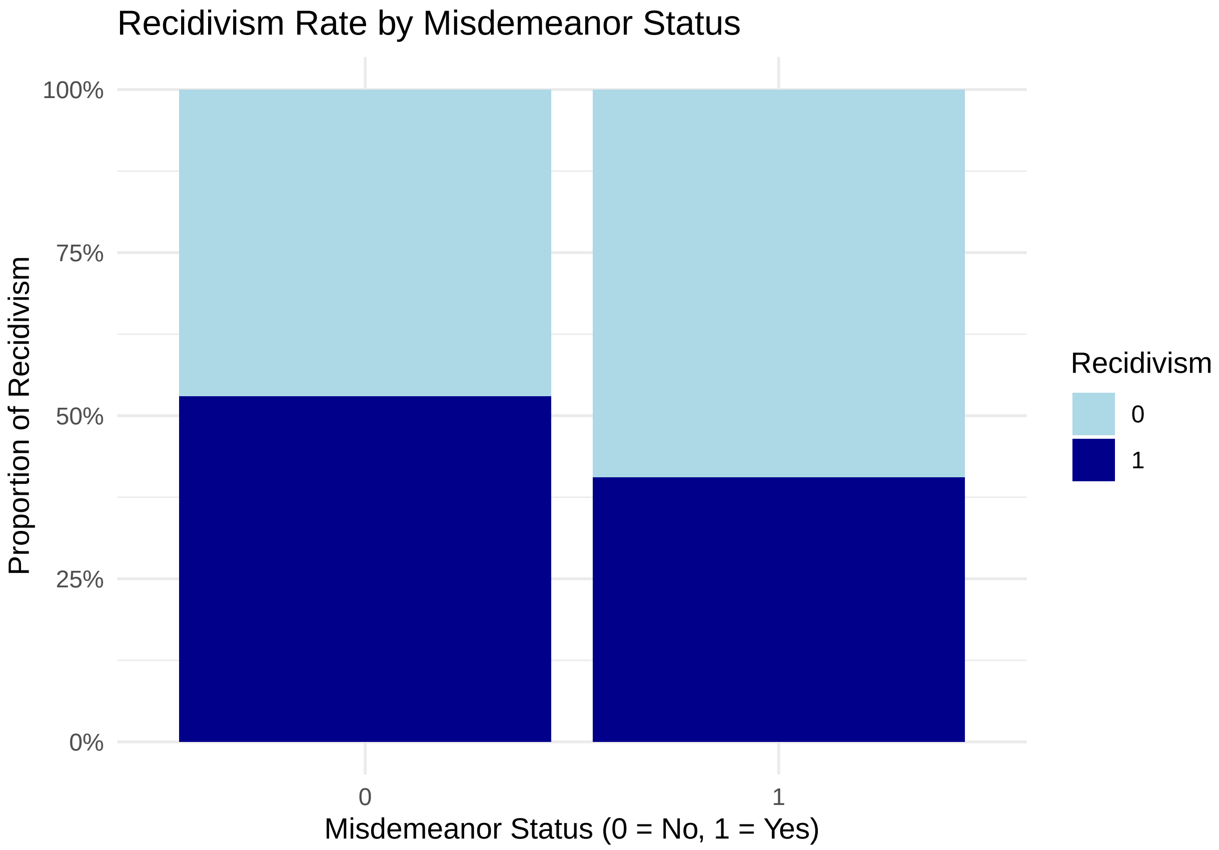


Let's dive deeper into the data, and see how misdemeanor and priors are related to recidivism.

```r
ggplot(dataset, aes(x = as.factor(recid), y = priors, fill = as.factor(recid))) +
  geom_boxplot() +
  labs(x = "Recidivism (0 = No, 1 = Yes)", y = "Number of Prior Convictions",
       fill = "Recidivism") +
  ggtitle("Distribution of Prior Convictions by Recidivism Status") +
  theme_minimal() +
  scale_fill_manual(values = c("0" = "lightblue", "1" = "darkblue"))
```



```r
ggplot(dataset, aes(x = as.factor(misdemeanour), fill = as.factor(recid))) +
  geom_bar(position = "fill") +
  labs(x = "Misdemeanor Status (0 = No, 1 = Yes)", y = "Proportion of Recidivism",
       fill = "Recidivism") +
  scale_y_continuous(labels = scales::percent) +
  ggtitle("Recidivism Rate by Misdemeanor Status") +
  theme_minimal() +
  scale_fill_manual(values = c("0" = "lightblue", "1" = "darkblue"))
```



It seems that individuals with more prior convictions are more likely to recidivate, the same applies to individuals convicted of felonies.

Now, we apply a logistic model, and see what we can conclude.

```
logit_model <- glm(recid ~ female + young + old + priors + misdemeanour +
                  af_am + asian + hisp + nat_am + other,data = dataset,
                  family = binomial)

summary(logit_model)
```

```
##
## Call:
## glm(formula = recid ~ female + young + old + priors + misdemeanour +
##     af_am + asian + hisp + nat_am + other, family = binomial,
##     data = dataset)
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.756673   0.128208  -5.902 3.59e-09 ***
## female       -0.516661   0.092418  -5.591 2.26e-08 ***
## young         0.841934   0.082492  10.206  < 2e-16 ***
## old          -0.638495   0.102060  -6.256 3.95e-10 ***
## priors        0.158902   0.009184  17.302  < 2e-16 ***
## misdemeanour -0.225267   0.073369  -3.070  0.00214 **
## af_am         0.264074   0.126018   2.096  0.03612 *
## asian        -0.415433   0.442353  -0.939  0.34766
## hisp         -0.014971   0.153951  -0.097  0.92253
## nat_am       -0.084262   0.657012  -0.128  0.89795
## other                 NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5638.6  on 4068  degrees of freedom
## Residual deviance: 4988.4  on 4059  degrees of freedom
## AIC: 5008.4
##
## Number of Fisher Scoring iterations: 4
```

Gender: The negative coefficient (-0.517) of female suggests that females are less likely to recidivate ($p < 0.001$).

Age: Young: The positive coefficient (0.842) indicates that younger individuals (under 25) are significantly more likely to recidivate ($p < 0.001$). Old: The negative coefficient (-0.639) indicates that older individuals (over 45) are significantly less likely to recidivate ($p < 0.001$).

Priors: The positive coefficient (0.159) suggests that each additional prior conviction increases the likelihood of recidivism ($p < 0.001$).

Misdemeanor: The negative coefficient (-0.225) suggests that individuals who committed a misdemeanor are less likely to recidivate compared to those who committed more serious crimes ($p = 0.002$).

Race: The positive coefficient (0.264) indicates that African American individuals have a higher likelihood to recidivate ($p = 0.036$).

The remaining variables are not statistically significant in this model.

Next, we simplify the model, and test it.

```r
# Generate training and testing sets
set.seed(42)

# 70% training, 30% testing
split <- sample.split(dataset_combined$recid, SplitRatio = 0.7)
training_set <- subset(dataset_combined, split == TRUE)
testing_set <- subset(dataset_combined, split == FALSE)

# Remove non-significant variables
logit_model_simplified <- glm(recid ~ female + young + old + priors +
                              misdemeanour + af_am,
                              family = binomial, data = training_set)

# Try logistic again
logit_model_simplified <- glm(recid ~ female + young + old + priors +
                              misdemeanour + af_am,
                              family = binomial, data = training_set)

# Summary
summary(logit_model_simplified)
```

```
##
## Call:
## glm(formula = recid ~ female + young + old + priors + misdemeanour +
##     af_am, family = binomial, data = training_set)
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -0.80135    0.10709  -7.483 7.25e-14 ***
## female       -0.57199    0.11139  -5.135 2.82e-07 ***
## young         0.92568    0.09939   9.314  < 2e-16 ***
## old          -0.58767    0.12125  -4.847 1.25e-06 ***
## priors        0.15790    0.01096  14.405  < 2e-16 ***
## misdemeanour -0.23710    0.08791  -2.697  0.00699 ** 
## af_am         0.30302    0.10194   2.973  0.00295 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 3946.6  on 2847  degrees of freedom
## Residual deviance: 3475.8  on 2841  degrees of freedom
## AIC: 3489.8
## 
## Number of Fisher Scoring iterations: 4
```

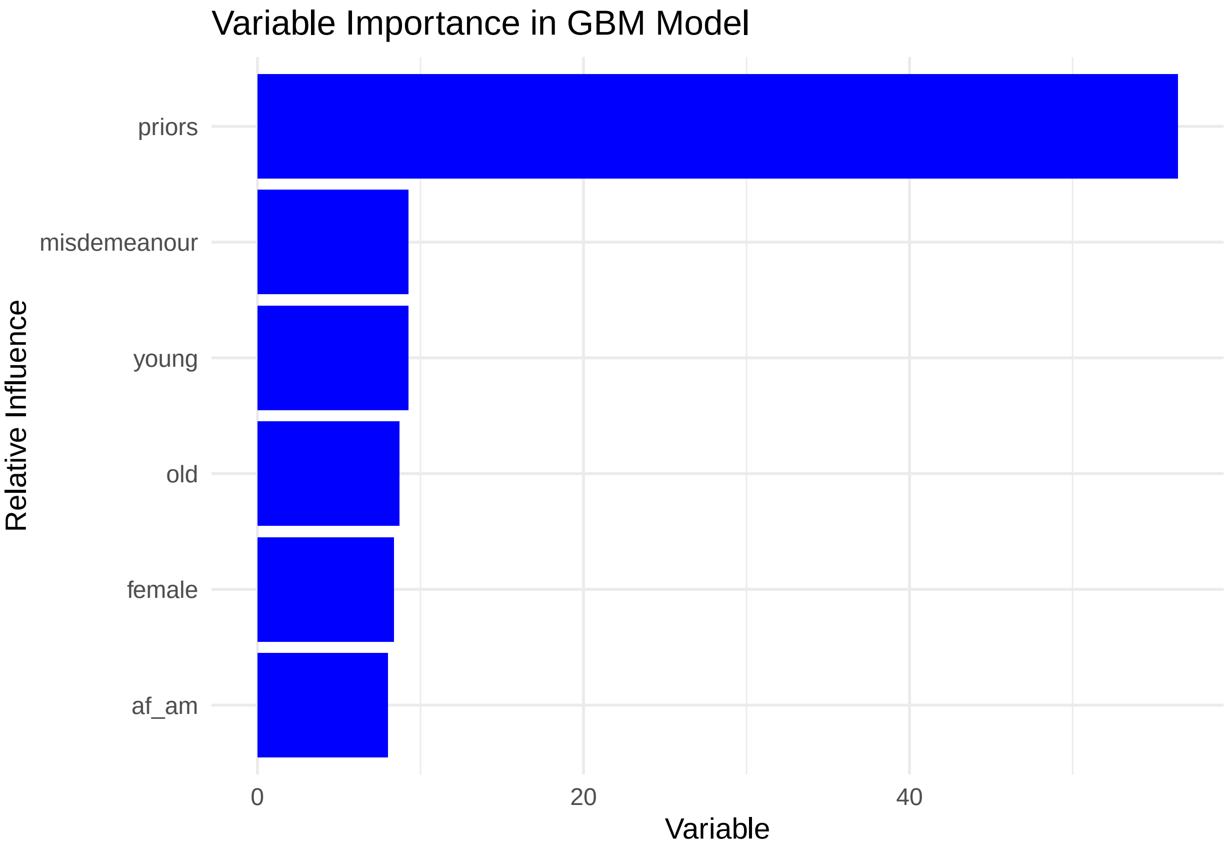
```
# Predict
predicted_probs_simplified <- predict(logit_model_simplified,
                                       newdata = testing_set, type = "response")
# using 0.5 as the threshold, may be arbitrary?
predicted_classes_simplified <- ifelse(predicted_probs_simplified > 0.5, 1, 0)

# Evaluate
confusion_matrix_simplified <- confusionMatrix(as.factor
                                               (predicted_classes_simplified),
                                               as.factor(testing_set$recid))

print(confusion_matrix_simplified)
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 423 221
##          1 202 375
##                                           
##                Accuracy : 0.6536          
##                  95% CI : (0.6261, 0.6803)
##     No Information Rate : 0.5119          
##     P-Value [Acc > NIR] : <2e-16          
##                                           
##                   Kappa : 0.3062          
##                                           
##  Mcnemar's Test P-Value : 0.3815          
##                                           
##             Sensitivity : 0.6768          
##             Specificity : 0.6292          
##          Pos Pred Value : 0.6568          
##          Neg Pred Value : 0.6499          
##              Prevalence : 0.5119          
##          Detection Rate : 0.3464          
##    Detection Prevalence : 0.5274
##       Balanced Accuracy : 0.6530
##
##        'Positive' Class : 0
##
```

The model seems to be moderately effective at predicting recidivism, with an accuracy of 65%. Let's see what we can get with other models.

GBM:

```r
# Generate model
set.seed(42)
gbm_model <- gbm(
  recid ~ female + young + old + priors + misdemeanour + af_am,
  data = training_set,
  distribution = "bernoulli",
  n.trees = 3000,
  interaction.depth = 5,
  shrinkage = 0.05, #sample size is relatively small
  cv.folds = 5
)

# Plot
summary(gbm_model, plot=FALSE) %>%
  as.data.frame() %>%
  ggplot(aes(x = reorder(var, rel.inf), y = rel.inf)) +
  geom_bar(stat = "identity", fill = "blue") +
  coord_flip() +  # Flip the coordinates to make labels horizontal
  labs(x = "Relative Influence", y = "Variable",
       title = "Variable Importance in GBM Model") +
  theme_minimal() +
  theme(axis.text.y = element_text(hjust = 1))
```



```r
# Predict on the testing set
predicted_probs_gbm <- predict(gbm_model, newdata = testing_set, n.trees =
                                 3000, type = "response")

# Convert probabilities to binary (using 0.5 as the threshold)
predicted_classes_gbm <- ifelse(predicted_probs_gbm > 0.5, 1, 0)

# Evaluate
confusion_matrix_gbm <- confusionMatrix(as.factor(predicted_classes_gbm),
                                        as.factor(testing_set$recid))

print(confusion_matrix_gbm)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 413 211
##          1 212 385
##
##                Accuracy : 0.6536
##                  95% CI : (0.6261, 0.6803)
##     No Information Rate : 0.5119
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.3068
##
```

```
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.6608
##             Specificity : 0.6460
##          Pos Pred Value : 0.6619
##          Neg Pred Value : 0.6449
##              Prevalence : 0.5119
##          Detection Rate : 0.3382
##    Detection Prevalence : 0.5111
##       Balanced Accuracy : 0.6534
##
##        'Positive' Class : 0
##
```

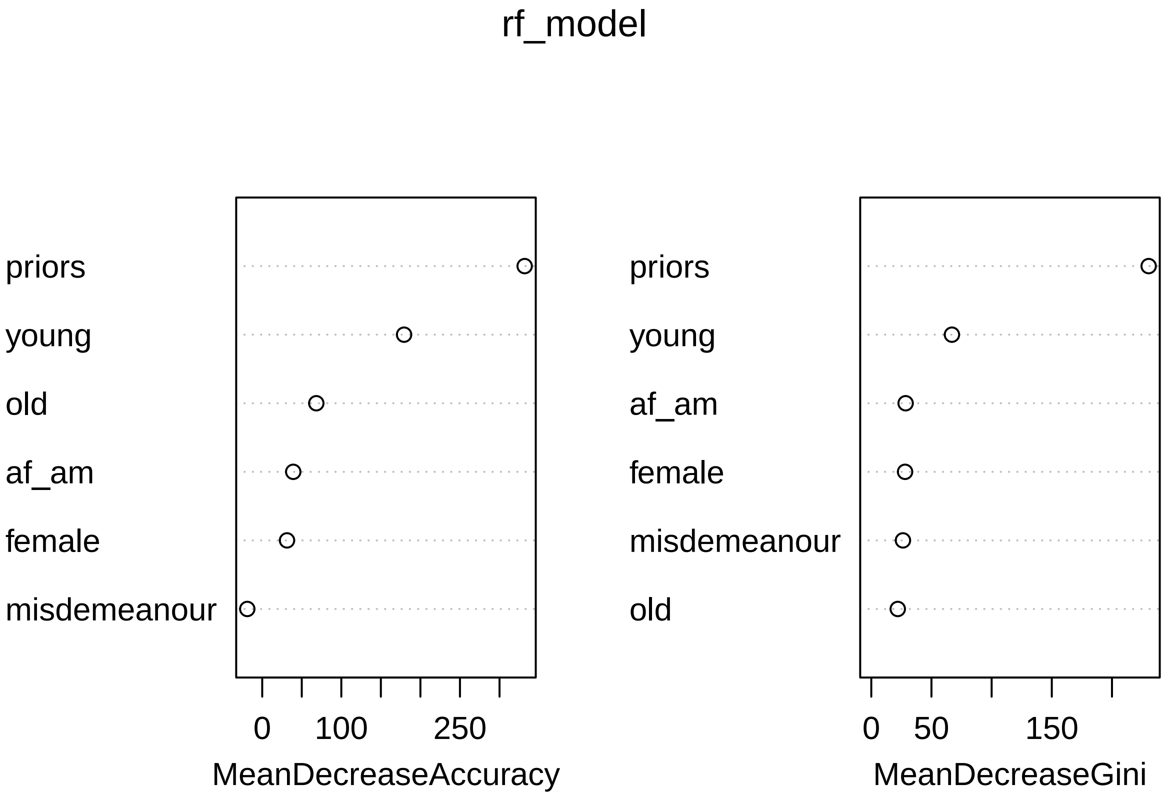
Accuracy is about 65%, and the plot shows that priors (number of prior conviction) is significantly more influential than other variables.

Random forest:

```r
# Ensure recid as factor for RF
training_set$recid <- as.factor(training_set$recid)
testing_set$recid <- as.factor(testing_set$recid)

# Train a Random Forest model for classification
set.seed(42)
rf_model <- randomForest(recid ~ female + young + old + priors +
            misdemeanour + af_am, data = training_set, ntree = 3000,
            mtry = 5, importance = TRUE)

# Predict on the testing set using the Random Forest model
predicted_rf <- predict(rf_model, newdata = testing_set)

# Evaluate the model's performance using the testing set
confusion_matrix_rf <- confusionMatrix(predicted_rf, testing_set$recid)

# Print the confusion matrix
print(confusion_matrix_rf)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 415 208
##          1 210 388
##
##                Accuracy : 0.6577
##                  95% CI : (0.6303, 0.6843)
##     No Information Rate : 0.5119
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.315
##
##  Mcnemar's Test P-Value : 0.961
##
##             Sensitivity : 0.6640
##             Specificity : 0.6510
##          Pos Pred Value : 0.6661
##          Neg Pred Value : 0.6488
##              Prevalence : 0.5119
##          Detection Rate : 0.3399
##    Detection Prevalence : 0.5102
##       Balanced Accuracy : 0.6575
##
##        'Positive' Class : 0
##
```

```
# Plot the importance of variables
varImpPlot(rf_model)
```



The accuracy is about 66%. Both plots suggest that priors (number of prior convictions) is the most important predictor in this model. This variable has the most significant impact on both the accuracy of the model and the purity of the splits in the decision trees. Other variables also have some importance but are less influential.

Overall, the three models demonstrated highly similar patterns and accuracy, with all achieving an accuracy between 65% and 66%. While this level of accuracy is acceptable, it falls short of being entirely satisfactory. Further tuning and optimization will be conducted in future work to improve model performance.