Assessing and Mitigating Algorithmic Bias in Criminal Risk Scores

Lucius Bynum, Preethi Seshadri

19 November 2017

Our Logistic Regression Model

Variable Selection

```
compas_data <- read.csv('compas-scores-two-years.csv')
```

Clean the compas dataset using the same approach Propublica uses.

```
compas_data <- compas_data %>%
  select(age, c_charge_degree, race, age_cat, score_text, sex, priors_count, days_b_screening_arrest, de
  filter(days_b_screening_arrest <= 30) %>%
  filter(days_b_screening_arrest >= -30) %>%
  filter(is_recid != -1) %>%
  filter(c_charge_degree != "O") %>%
  filter(score_text != 'N/A')
```

We will include the time spent in jail (scaled to be in days).

```
compas_data$c_jail_out <- strptime(compas_data$c_jail_out, format = '%Y-%m-%d %H:%M:%S')
compas_data$c_jail_in <- strptime(compas_data$c_jail_in, format = '%Y-%m-%d %H:%M:%S')
compas_data$time_spent <- difftime(compas_data$c_jail_out, compas_data$c_jail_in, units='days')
compas_data$time_spent <- as.numeric(compas_data$time_spent)
compas_data <- compas_data[, !(colnames(compas_data) %in% c("c_jail_out", "c_jail_in"))]
```

We will remove examples that do not have a time spent in jail value.

```
compas_data <- compas_data[!is.na(compas_data$time_spent),]
```

```
# perform k-fold cross validation
compas_data <- compas_data[sample(1:nrow(compas_data)), ] # shuffle the rows up
folds <- cut(seq(1,nrow(compas_data)),breaks=10,labels=FALSE) # divide into 10 folds
#Perform 10 fold cross validation
matrices <- list()
race_bias <- c()
sex_bias <- c()
age_bias <- c()
for(i in 1:10) { # loop through each fold and fit model
  vars <- compas_data %>% select(sex, age, ends_with('count'), c_charge_degree, time_spent, is_recid)
  indices <- which(folds==i,arr.ind=TRUE)
  testData <- vars[indices, ]
  trainData <- vars[-indices, ]
  compas_test <- compas_data[indices, ]

  # fit a logistic regression model on training data to predict recidivism
  model <- glm(is_recid ~., family = binomial(link='logit'), data = trainData)

  # compute predictions and confusion matrix
```

```
  preds <- predict(model, newdata = testData, type = "response") # response gives probability instead o
  matrices[[i]] <- confusionMatrix(data = as.numeric(preds>0.5), reference = testData$is_recid)

  # compute low, medium, high scores
  our_scores <- compas_test %>%
  mutate(new_score = preds) %>%
  mutate(new_score_text = as.factor(
      ifelse(new_score >= 0.75, 'High',
             ifelse(new_score >= 0.45, 'Medium', 'Low')))
  )

  # create a new dataframe for predicting probability of recidivating (as specified by our model)
  new_df <- our_scores %>%
    mutate(crime_factor = factor(c_charge_degree)) %>%
    mutate(age_factor = as.factor(age_cat)) %>%
    within(age_factor <- relevel(age_factor, ref = 1)) %>%
    mutate(race_factor = factor(race)) %>%
    within(race_factor <- relevel(race_factor, ref = 3)) %>%
    mutate(gender_factor = factor(sex, labels= c("Female","Male"))) %>%
    within(gender_factor <- relevel(gender_factor, ref = 2)) %>%
    mutate(score_factor = factor(new_score_text != "Low", labels = c("LowScore","HighScore")))

  # new logistic regression model to predict our computed recidivism probabilities (on the test data)
  new_pp_model <- glm(score_factor ~ gender_factor + age_factor + race_factor + priors_count + crime_fa

  model_intercept <- coef(new_pp_model)['(Intercept)']
  # Bias race
  black_coef <- coef(new_pp_model)['race_factorAfrican-American']
  race_bias[i] <- (sigmoid(model_intercept + black_coef) / sigmoid(model_intercept))

  # Bias sex
  woman_coef <- coef(new_pp_model)['gender_factorFemale']
  sex_bias[i] <- (sigmoid(model_intercept + woman_coef) / sigmoid(model_intercept))

  # Bias age
  age_coef <- coef(new_pp_model)['age_factorLess than 25']
  age_bias[i] <- (sigmoid(model_intercept + age_coef) / sigmoid(model_intercept))
}
```

```
matrices
```

```
## [[1]]
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 254 109
##          1  81 174
##
##                Accuracy : 0.693
##                  95% CI : (0.655, 0.729)
##     No Information Rate : 0.542
##     P-Value [Acc > NIR] : 1.58e-14
##
```

```
##                  Kappa : 0.376
##  Mcnemar's Test P-Value : 0.0501
##
##            Sensitivity : 0.758
##            Specificity : 0.615
##         Pos Pred Value : 0.700
##         Neg Pred Value : 0.682
##             Prevalence : 0.542
##         Detection Rate : 0.411
##   Detection Prevalence : 0.587
##      Balanced Accuracy : 0.687
##
##       'Positive' Class : 0
##
##
## [[2]]
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 209 124
##          1  87 197
##
##               Accuracy : 0.658
##                 95% CI : (0.619, 0.695)
##    No Information Rate : 0.52
##    P-Value [Acc > NIR] : 3.03e-12
##
##                  Kappa : 0.318
##  Mcnemar's Test P-Value : 0.0132
##
##            Sensitivity : 0.706
##            Specificity : 0.614
##         Pos Pred Value : 0.628
##         Neg Pred Value : 0.694
##             Prevalence : 0.480
##         Detection Rate : 0.339
##   Detection Prevalence : 0.540
##      Balanced Accuracy : 0.660
##
##       'Positive' Class : 0
##
##
## [[3]]
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 231 111
##          1  93 182
##
##               Accuracy : 0.669
##                 95% CI : (0.631, 0.706)
##    No Information Rate : 0.525
```

```
##     P-Value [Acc > NIR] : 2.6e-13
##
##                   Kappa : 0.335
##  Mcnemar's Test P-Value : 0.234
##
##             Sensitivity : 0.713
##             Specificity : 0.621
##          Pos Pred Value : 0.675
##          Neg Pred Value : 0.662
##              Prevalence : 0.525
##          Detection Rate : 0.374
##    Detection Prevalence : 0.554
##       Balanced Accuracy : 0.667
##
##        'Positive' Class : 0
##
##
## [[4]]
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 223  95
##          1  94 205
##
##                Accuracy : 0.694
##                  95% CI : (0.656, 0.73)
##     No Information Rate : 0.514
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.387
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.703
##             Specificity : 0.683
##          Pos Pred Value : 0.701
##          Neg Pred Value : 0.686
##              Prevalence : 0.514
##          Detection Rate : 0.361
##    Detection Prevalence : 0.515
##       Balanced Accuracy : 0.693
##
##        'Positive' Class : 0
##
##
## [[5]]
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 256 104
##          1  75 182
##
##                Accuracy : 0.71
```

```
##                  95% CI : (0.672, 0.745)
##     No Information Rate : 0.536
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.413
##  Mcnemar's Test P-Value : 0.0364
##
##             Sensitivity : 0.773
##             Specificity : 0.636
##          Pos Pred Value : 0.711
##          Neg Pred Value : 0.708
##              Prevalence : 0.536
##          Detection Rate : 0.415
##    Detection Prevalence : 0.583
##       Balanced Accuracy : 0.705
##
##        'Positive' Class : 0
##
##
## [[6]]
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 232 107
##          1  84 194
##
##                Accuracy : 0.69
##                  95% CI : (0.652, 0.727)
##     No Information Rate : 0.512
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.379
##  Mcnemar's Test P-Value : 0.111
##
##             Sensitivity : 0.734
##             Specificity : 0.645
##          Pos Pred Value : 0.684
##          Neg Pred Value : 0.698
##              Prevalence : 0.512
##          Detection Rate : 0.376
##    Detection Prevalence : 0.549
##       Balanced Accuracy : 0.689
##
##        'Positive' Class : 0
##
##
## [[7]]
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 213 116
##          1  90 198
```

```
##
##                Accuracy : 0.666
##                  95% CI : (0.627, 0.703)
##     No Information Rate : 0.509
##     P-Value [Acc > NIR] : 2.07e-15
##
##                   Kappa : 0.333
##  Mcnemar's Test P-Value : 0.0815
##
##             Sensitivity : 0.703
##             Specificity : 0.631
##          Pos Pred Value : 0.647
##          Neg Pred Value : 0.688
##              Prevalence : 0.491
##          Detection Rate : 0.345
##    Detection Prevalence : 0.533
##       Balanced Accuracy : 0.667
##
##        'Positive' Class : 0
##
##
## [[8]]
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 227 103
##          1  93 194
##
##                Accuracy : 0.682
##                  95% CI : (0.644, 0.719)
##     No Information Rate : 0.519
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.363
##  Mcnemar's Test P-Value : 0.52
##
##             Sensitivity : 0.709
##             Specificity : 0.653
##          Pos Pred Value : 0.688
##          Neg Pred Value : 0.676
##              Prevalence : 0.519
##          Detection Rate : 0.368
##    Detection Prevalence : 0.535
##       Balanced Accuracy : 0.681
##
##        'Positive' Class : 0
##
##
## [[9]]
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
```

```
##          0 223 117
##          1  88 189
##
##                Accuracy : 0.668
##                  95% CI : (0.629, 0.705)
##     No Information Rate : 0.504
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.335
##  Mcnemar's Test P-Value : 0.0505
##
##             Sensitivity : 0.717
##             Specificity : 0.618
##          Pos Pred Value : 0.656
##          Neg Pred Value : 0.682
##              Prevalence : 0.504
##          Detection Rate : 0.361
##    Detection Prevalence : 0.551
##       Balanced Accuracy : 0.667
##
##        'Positive' Class : 0
##
##
## [[10]]
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 236  99
##          1  93 190
##
##                Accuracy : 0.689
##                  95% CI : (0.651, 0.726)
##     No Information Rate : 0.532
##     P-Value [Acc > NIR] : 1.39e-15
##
##                   Kappa : 0.375
##  Mcnemar's Test P-Value : 0.718
##
##             Sensitivity : 0.717
##             Specificity : 0.657
##          Pos Pred Value : 0.704
##          Neg Pred Value : 0.671
##              Prevalence : 0.532
##          Detection Rate : 0.382
##    Detection Prevalence : 0.542
##       Balanced Accuracy : 0.687
##
##        'Positive' Class : 0
##
```

```
race_bias
```

```
##  [1] 0.9984 0.7295 0.9957 1.3615 1.0152 1.2967 1.6990 1.1032 1.0141 1.2001
```

```
sex_bias
```

```
##  [1] 0.013169 0.010135 0.028070 0.019497 0.027228 0.021643 0.042199
##  [8] 0.022203 0.011552 0.006546
```

```
age_bias
```

```
##  [1] 3.028 7.414 4.865 4.500 4.514 4.729 5.778 3.912 8.377 3.917
```