Josiah Kim

Professor Zorn

PLSC 476

30 April 2021

Removing Racial Bias in Predicting Recidivism

Overview and Hypothesis

In class, we discussed ProPublica's analysis on the COMPAS recidivism algorithm[1]. It turns out that the simple logistic regression model we examined in the class practicum predicts recidivism well. Not only that, but a K-fold validation procedure verified that the class practicum model predicted recidivism better than the COMPAS model. More importantly, ProPublica found that there is a significant algorithmic bias in the COMPAS model with regards to the fact that black defendants were deemed as higher risk for recidivism than their white counterparts.

In hopes of decreasing the effects of algorithmic bias, I theorize that removing race as a predictor for recidivism will not significantly change the class practicum model's predictive capabilities for recidivism. While I do not expect the predictive capabilities for recidivism to improve without race as a predictor, I believe that there will be no statistically significant reason to continue using race as a predictor for recidivism.

Data and Measurement

The Cox Dataset [2] I will draw upon was compiled by ProPublica and used in the analysis of the COMPAS recidivism algorithm. The dataset contains 10,331 cases each of which represent previously incarcerated individuals. We are provided biographical information, previous history within the judicial system, COMPAS scores, and recidivism for each individual.

[1] Found at https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[2] Found at https://raw.githubusercontent.com/propublica/compas-analysis/master/cox-parsed.csv

My main outcome variable is recidivism (*Recid*). This variable takes binary values to signify if the individual committed recidivism (= *1*) or not (= *0*). My predictor variables include sex (*sex*), age (*age*), race (*race*), juvenile felonies (*JuvFelonies*), juvenile misdemeanors (*JuvMisdem*), prior arrests (*Priors*), and if there was a felony charge (*FelonyCharge*). My analysis will contain a comparison between the class practicum model for predicting recidivism (the original model) and a modified model for predicting recidivism without using race as a predictor variable.

Analysis and Findings

Table 1 shows the original model for predicting recidivism. The greatest positive predictor for recidivism that is also statistically significant is sex. A male offender is 38.6 percent more at risk for recidivism than female offenders.

Conversely, the greatest negative predictor for recidivism that is also the most statistically significant is if the offender's race is categorized as "other." An offender in the "other" race category is 30.7 percent less likely to commit recidivism. In fact, three race values are statistically significant for predicting recidivism.

Table 2 shows the modified model for predicting recidivism. We can see that sex is still the greatest positive predictor for recidivism that is also statistically significant. However, compared to the original model, the effect is slightly weaker (by 0.8 percent). Furthermore, the *priors* variable remains statistically significant in the modified model while the effect becomes slightly stronger (by 0.5 percent).

```
==============================================
                 Dependent variable:
               ---------------------------
                       Recid
----------------------------------------------
SexMale                 0.368***
                        (0.058)

Age                    -0.040***
                        (0.002)

RaceAsian                -0.407
                        (0.350)

RaceCaucasian            -0.111**
                        (0.050)

RaceHispanic             -0.303***
                        (0.083)

RaceNative American        -0.168
                        (0.398)

RaceOther                -0.307***
                        (0.105)

JuvFelonies               0.028
                        (0.050)

JuvMisdem                  0.068
                        (0.054)

Priors                  0.116***
                        (0.005)

FelonyCharge               0.018
                        (0.048)

Constant                 0.053
                        (0.093)

----------------------------------------------
Observations              10,331
Log Likelihood           -6,051.861
Akaike Inf. Crit.        12,127.720
==============================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

Table 1. Original Logistic Regression Model for Recidivism

```
============================================
                 Dependent variable:
               ---------------------------
                       Recid
--------------------------------------------
SexMale                 0.360***
                        (0.058)

Age                    -0.041***
                        (0.002)

JuvFelonies               0.032
                        (0.051)

JuvMisdem                  0.073
                        (0.054)

Priors                  0.121***
                        (0.005)

FelonyCharge               0.027
                        (0.048)

Constant                 -0.005
                        (0.092)

--------------------------------------------
Observations              10,331
Log Likelihood           -6,062.657
Akaike Inf. Crit.       12,139.310
============================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

Table 2. Modified Logistic Regression Model for Recidivism Without the Race Predictor

A confusion matrix for the modified model shows that the model correctly predicted 93.26 precent of true positive cases and 21.85 percent of true negative cases. This is comparable to the original model which correctly predicted 93.13 percent of true positive cases and 22.26 of true negative cases. The modified model correctly predicted 0.13 percent more true positive cases and 0.41 true negative cases.

Furthermore, the overall accuracies of the models are also fairly comparable. The original model has an accuracy of 69.31 percent compared to the modified model's accuracy of 69.26

percent. In general, we can see that the modified model is slightly less accurate than the original model (by 0.05 percent).

Figure 1 shows the comparison in ROC curves for both the original model and the modified model. The AUC for both the models are close to 0.7 with the original model being slightly higher (by 0.001) which may be a negligible in telling us which model has a better measure of separability.

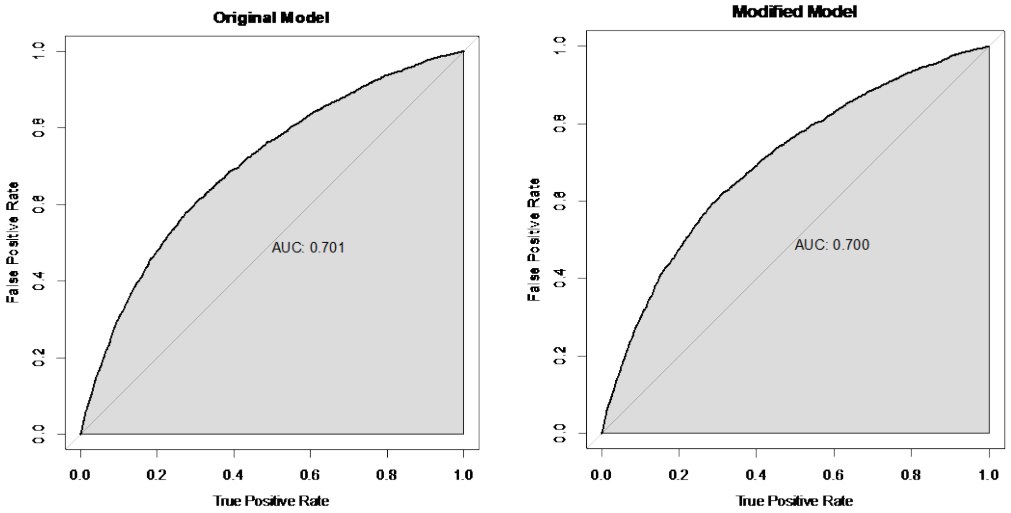

Figure 1. ROC Curves

These models perform similarly in-sample. However, figure 2 shows the performance of the models out-of-sample. Setting K=10, we can see that both models have the similar performance metrics when it comes to predicting recidivism out-of-sample. The AUC-ROC mean for both models is 0.699.

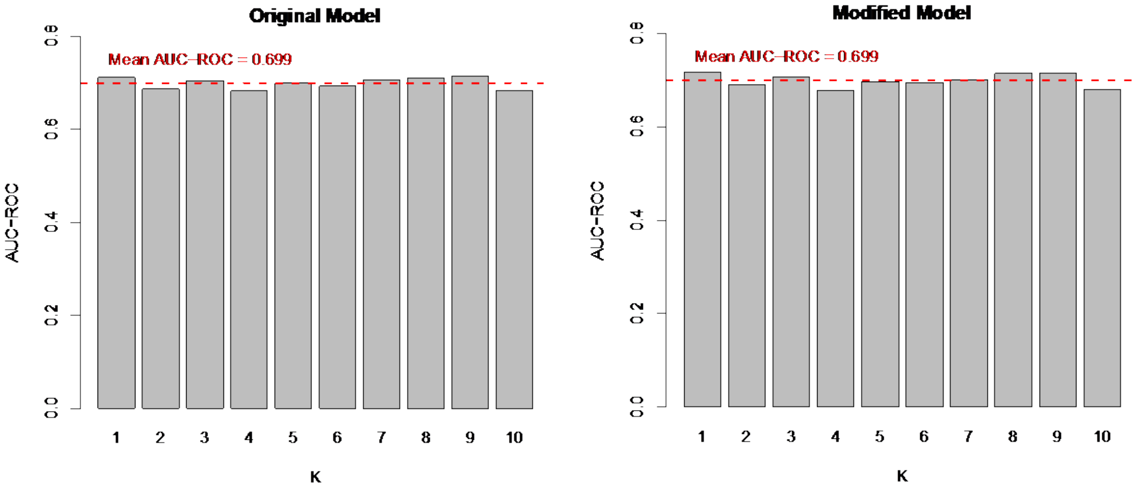

Figure 2. K-fold Cross-Validation

Summary

By comparing the original practicum model to the modified model without race as a predictor, we can see that the difference in performance is perhaps insignificant. Both models predict in-sample similarly with the modified model showing an accuracy of 0.05 percent less than the original model. However, out-of-sample both models perform almost identically. This research module, although not conclusive, shows evidence that removing race as a predictor for recidivism will not significantly change a model's capabilities for predicting recidivism.