Technical University of Munich

Elektrotechnik

Maschinelle Intelligenz und Gesellschaft (in Python)

Final Homework Project

Authors:
Seif Werghi 03698296
Oussama Sayari 03700794
Marwen Mokni 03727422
Firas Guediri 03690798
Moez Ben Ayed 03702600

August 30, 2020

Contents

1 Data Preprocessing and Analysis **2**

2 Groups suited for Fairness **2**

3 The new binary Classifier **5**

4 Classifier checking **6**
4.1 Race Fairness 6
4.2 Gender Fairness 8
4.3 Age Fairness 8

5 Final thoughts and conclusions **8**

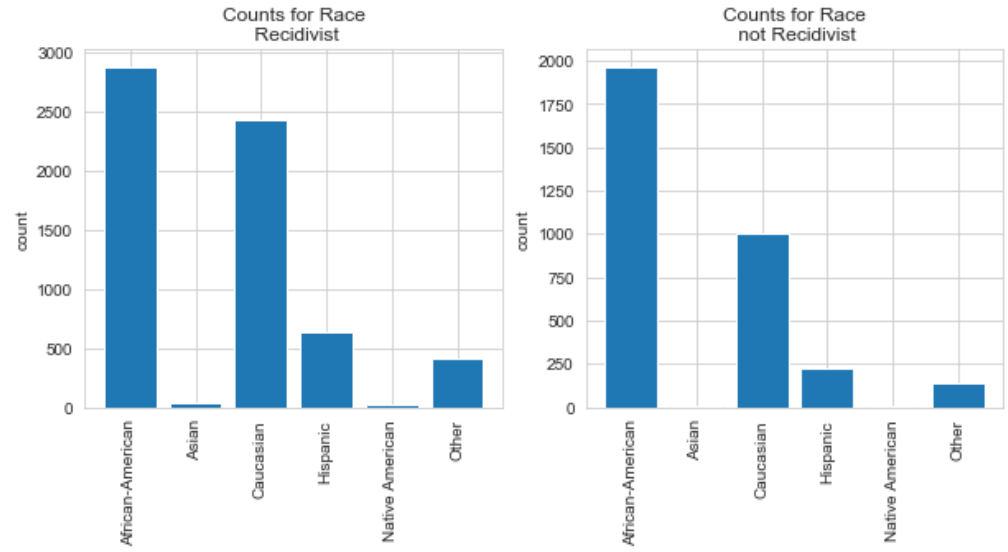
1 Data Preprocessing and Analysis

Out of the many features in the dataset[1], *Sex and Race* have caught our attention the most as we think they are mostly related to fairness. Individuals in society tend to use such definitions to make judgements about other individuals, decide whom to befriend and whom to discriminate. As a result, we try to catch a blink in figure 1 of the effect of these features on actual Recidivism and whether actually men commit more (or less) crimes compared to women.
The distribution does not tell much about the impact of these features on the recidivism. As a matter of fact, both cases look similar; *African-American having the highest count and Asian count being the lowest.* The same is applied for the Sex as females *displaying lesser count.* However, this might be due to the total count of each variable in the original data itself and the fact that not so many Native American or Asian or Females of that matter have been included in the study compared to the other majorities.
Age seem to have a more meaningful representation. In other words, young individuals varying in the interval $[18, 30]$ show the highest count of recidivism. It is also worth mentioning that African-American and Caucasian's count is more emphasised in the dataset. This is actually one of the reasons why Age has a big impact on the predictions.

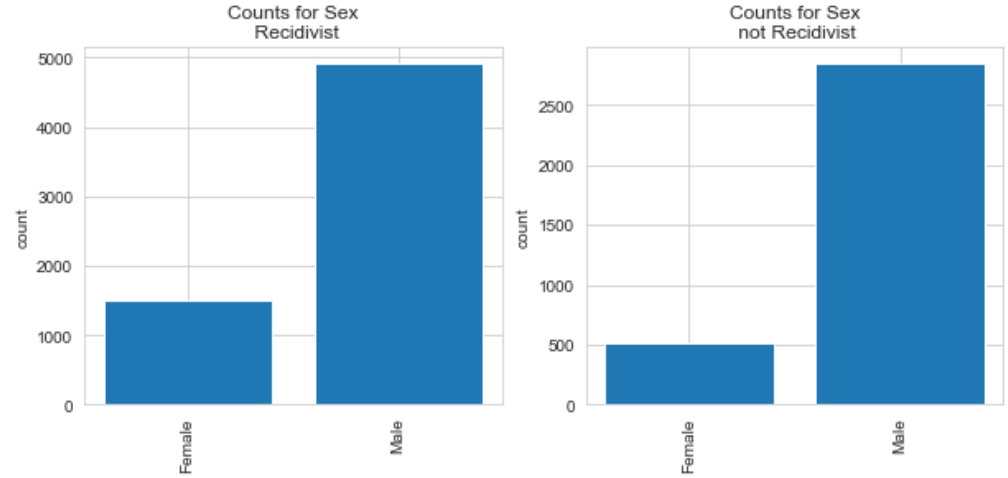
2 Groups suited for Fairness

Now, that we have a good understanding of how the data is distributed, let's see the effect of each group of race on Recidivism.
First, we tested the ROC Curves of every race using the *sklearn* library to actually understand the threshold. However, this method does not offer enough information. Moreover, races like Asian and Native American have such a low sample count that it was almost impossible to interpret the data (3 Native-American in the whole dataset). Incidentally, we are going to let go of these minorities for some of the studies to avoid miscalculations. As a consequence of this failure to achieve any logical interpretation we are forced to use another criteria to evaluate the performance and discrimination of the predictor. *"Most of the proposed fairness criteria are properties of the joint distribution of the sensitive attribute A, the target variable Y, and the*

[1]For a complete data overhaul, please refer to the Python Notebook.




```
Recidivism count according to the feature : Sex
```




```
Recidivism count according to the feature : Age
```

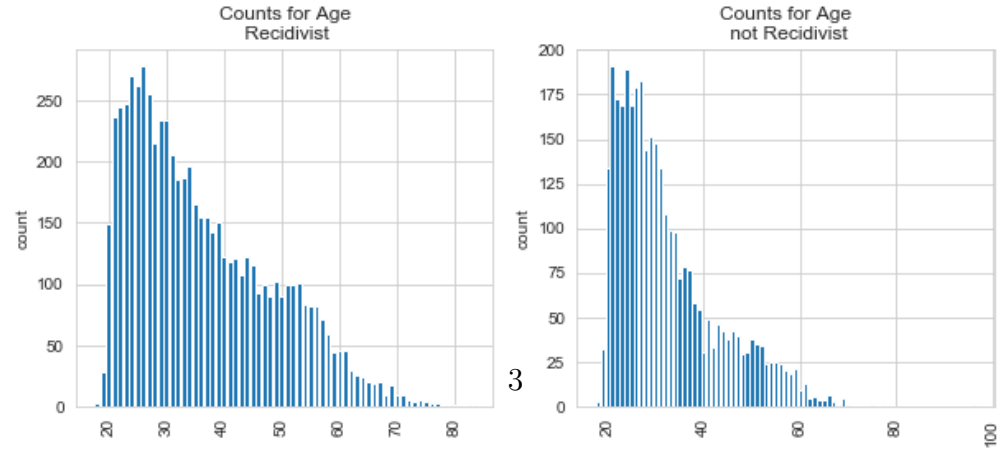


Figure 1: Recidivism distribution according to Race, Sex and Age.

classifier or score R." [2] According to the FairMLBook (Chap 2), these criteria fall into one of three different categories defined along the lines of different conditional statements: **Independence, Separation, Sufficiency**

Independence: The probability of being classified by the algorithm in each of the groups is equal for two individuals with different sensitive characteristics.

$$P(R = r|A = a) = P(R = r|A = b) \quad \forall r \in R \quad \forall a, b \in A$$

In the case of a binary classifier, this translates into:

$$\frac{TP_{A=a} + FP_{A=a}}{\#Group\ a} = \frac{TP_{B=b} + FP_{B=b}}{\#Group\ b} \tag{1}$$

Figure 2 summarises what we have done to understand the three terms. As African American and Caucasian have the major distribution, we decided to continue our studies based as such.
We have had an interesting conversation about the Independency. We have then reached the conclusion that the threshold actually does not have an effect on the term as (TP + FP) is always constant no matter what the threshold is. The same is applied to the count of Group (eg. African-American count). Additionally, we have introduced a relaxation term ϵ to give a small headroom for this equation (=0.05). Yet, the Independence was still never fulfilled. Which means an African-American actually does not have the same chances of receiving a score as a Caucasian individual. This can lead to a sort of discrimination in a worst case scenario.

Separation: The probability of being classified by the algorithm in each of the groups is equal for two individuals with different sensitive characteristics given that they actually belong in the same group (have the same target variable).

$$P(R = r|Y = q, A = a) = P(R = r|Y = q, A = b)$$
$$\forall r \in R \quad q \in Y \quad \forall a, b \in A$$

[2] Taken from Fairness and machine learning Limitations and Opportunities

In the case of a binary classifier, this translates into:

$$\frac{TP_{A=a}}{TP_{A=a} + FP_{A=a}} = \frac{TP_{A=b}}{TP_{A=b} + FP_{A=b}} \tag{2}$$

$$\frac{TN_{A=a}}{TN_{A=a} + FN_{A=a}} = \frac{TN_{A=b}}{TN_{A=b} + FN_{A=b}} \tag{3}$$

Sufficiency: The probability of actually being in each of the groups is equal for two individuals with different sensitive characteristics given that they were predicted to belong to the same group.

$$P(Y = q|R = r, A = a) = P(Y = q|R = r, A = b)$$
$$\forall q \in Y \quad r \in R \quad \forall a, b \in A$$

In the case of a binary classifier, this translates into:

$$\frac{TP_{A=a}}{TP_{A=a} + FN_{A=a}} = \frac{TP_{A=b}}{TP_{A=b} + FN_{A=b}} \tag{4}$$

$$\frac{FP_{A=a}}{FP_{A=a} + FN_{A=a}} = \frac{FP_{A=b}}{FP_{A=b} + FN_{A=b}} \tag{5}$$

Figure 2 also illustrates that **Separation and Sufficiency** is hard to achieve for most of thresholds. Which means that it is almost impossible to get a fair model between these two races. Maybe it could have been better if we had more samples for the other groups as well, or if we genuinely used another feature to split the data intro groups (Sex or Age).

3 The new binary Classifier

For our project we have chosen a Decision Tree as a binary classifier, as it was easy to realize and to understand. Tree Binary classifiers have shown great leaps and success in such situations so it should also yield great results for our case. The classifier tried to classify our groups according to their age as well as their previous convictions. We eliminated in a first time the race then the gender and lastly the age and checked if these three attributes do influence the fairness of our classifier or not.

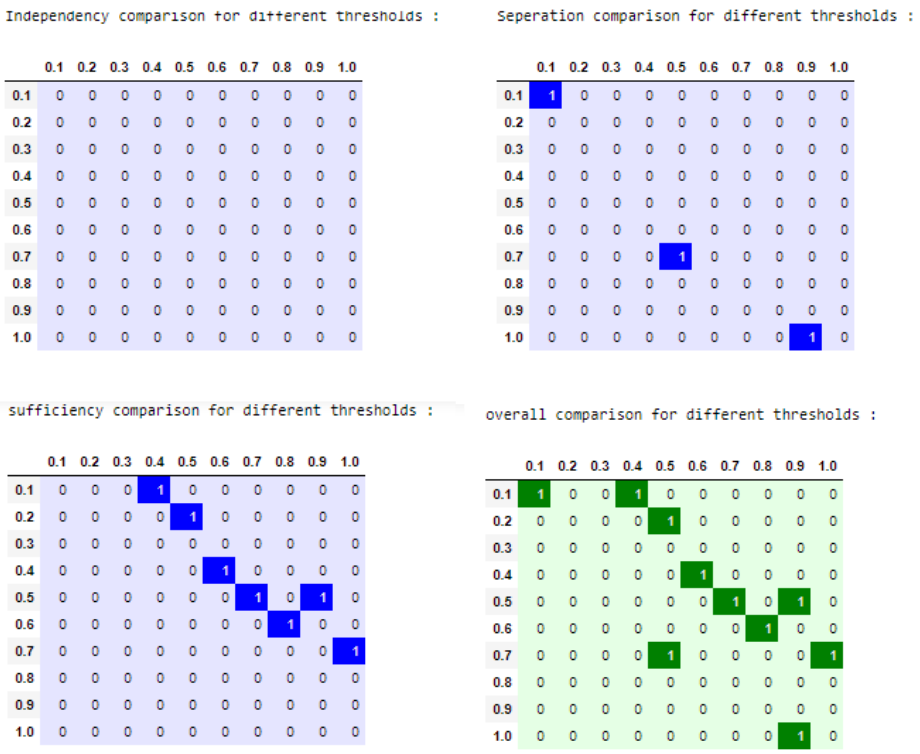
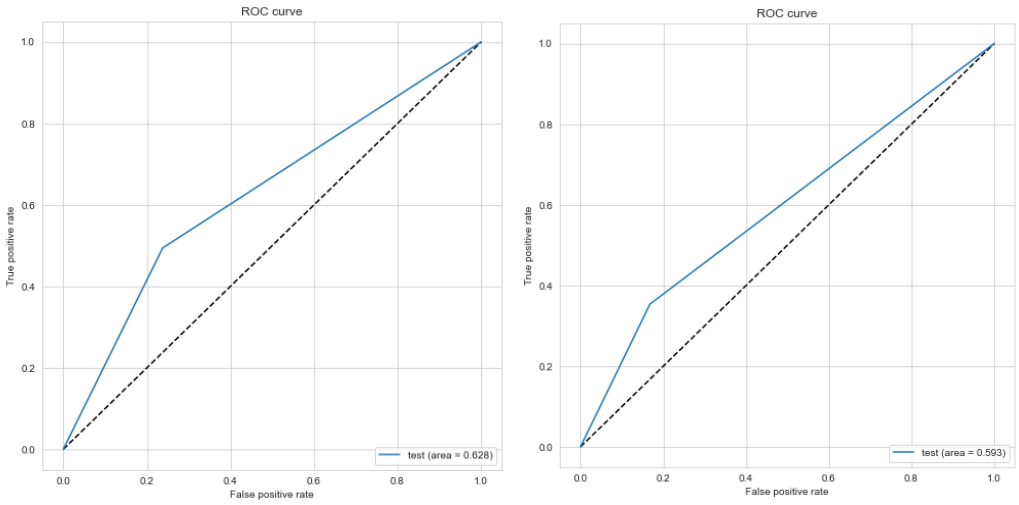


Figure 2: Independency, Sufficiency ,and Separation according to different thresholds for the two groups : African American and Caucasian; 0 Not fulfilled, 1 fulfilled

4 Classifier checking

4.1 Race Fairness

After excluding the Race feature from the prediction, it has yielded an accuracy of ≈ 0.67 (Which is almost equal to the COMPAS Score). We use the tree to predict solely on an African American subdataset which results in an a ROC Curve worse than the original given model. The same is also applied for the Caucasians, explained in figure 3. However, the Fairness Criteria (mentioned in Section 2) is yet still not fulfilled. Interestingly, African American show higher rates than whites in the three different criteria as shown in Figure 4



(a) African American ROC Curve (b) Caucasian Roc Curve

Figure 3: Race ROC Curves of the Tree Classifier(Test only)

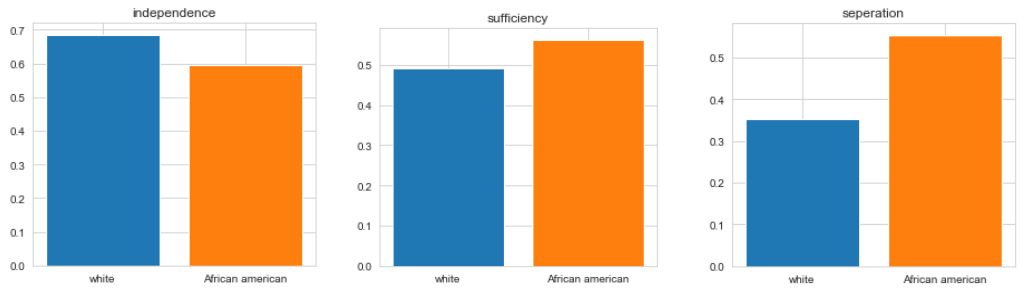


Figure 4: Conditional Criteria for the RACE feature in the Tree

4.2 Gender Fairness

After eliminating the Gender feature we notice that the accuracy is identical to that where we eliminated the race feature. This reveals to us that race but also gender don't have a big influence on the accuracy of our system. However, we notice from the ROC curve that men are discriminated against women because they only represent a precision of ≈ 0.30 in comparison to a ≈ 0.49 precision rate for women. It's also worth mentioning that even if the males have higher accuracy than females. The model actually works better for females and seems closer to the actual COMPAS Predictor, making it an unfair representation.

4.3 Age Fairness

In the age feature, we observe that the ROC curve for the groups under 25 is approximately a line at 1:1 which doesn't help us predict the recidivism for that group of people. In other words, the system is behaving randomly and spitting random results with a ≈ 0.52 accuracy. This has been first portrayed in the Decision Tree, when we were studying the other features and the predictor simply ignored these features and focused on Age as it had bigger weight. And then confirmed when we excluded age and the classifier got immensely worse. Going even further, the Fairness Criteria have been once again labeled as : not fulfilled.

5 Final thoughts and conclusions

We had first insights that actually not including Race and Gender in the classifier would result in worse results (≤ 0.5). However, it appears that Race and Gender aren't that much important. As a matter of fact, Age has shown that it plays a bigger role in this prediction. Nonetheless, we also need to admit: the dataset has in fact unfair distribution and collection of information and we had to only focus on some of the races, we choose a Tree classifier too that should find "its own good case" without hustle, using a different binary classifier (like a linear regression and then applying our own thresholds) might introduce other problems or worse results. Finally, there are other fairness criteria that we could be missing, such as Statistical Parity, Equalized odds, calibration etc...