IN SEARCH FOR FAIRNESS IN MACHINE LEARNING MODEL : COMPAS DATASET

- Manuel **SULMONT**
- Nairit **BANDYOPADHYAY**
- Archit **YADAV**

PREMISE

We are going to analyze the COMPAS algorithm.

This algorithm is used to assign certain risk factors (low, medium, high) to the criminals based on the questionnaire answered by the defendants.

The COMPAS is feed with these data and it output the “score-text”.

TARGETS : HIGH LEVEL VIEW

We aim to do the following:

To clean the dataset and perform some basic data exploration.

To use some general classifier on the dataset.

To check if there is bias.

To create a fair classifier.

PREPROCESSING

We followed the guidelines specified in propublica notebook.

Basically we selected some specific records that will be useful for our analysis, instead of choosing all the records.

EXPLORING THE DATASET

Explorations like trying to see like the decile scores of each race, number of defendants of each race and other interesting measures.

```
People got re-arrested by race
race
African-American    1661
Asian                  8
Caucasian            822
Hispanic             189
Native American        5
Other                124
Name: two_year_recid, dtype: int64
```

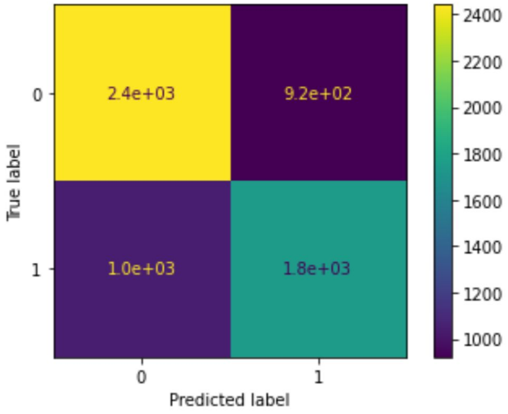
```
People got different categories of compas scores based on race
score_text  race
High        African-American     845
            Asian                  3
            Caucasian            223
            Hispanic              47
            Native American        4
            Other                 22
Low         African-American    1346
            Asian                 24
            Caucasian           1407
            Hispanic             368
            Native American        3
            Other                273
Medium      African-American     984
            Asian                  4
            Caucasian            473
            Hispanic              94
            Native American        4
            Other                 48
Name: race, dtype: int64
```

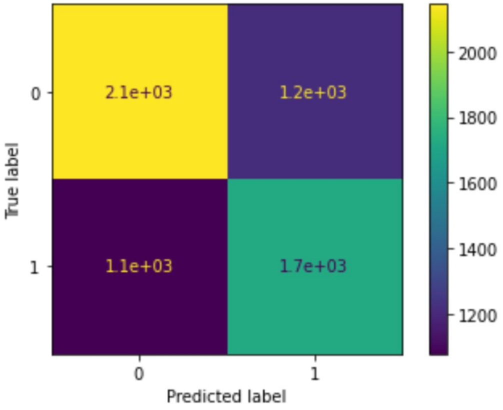
1829

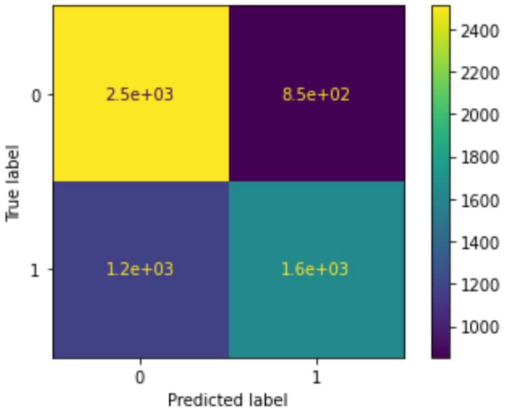
696

USING BASIC CLASSIFIERS

- We tried to **predict** whether the defendant **recidivated in 2 years** using some general classifiers
  - MLP
  - GaussianNB
  - SVM
- The accuracy of this classifiers **ranged around 60 - 67%**
- All of them had **high number of false positives (FP)** as shown in the confusion matrix







IS THERE A BIAS ?

- We created a **logistic regression model** and tried to mimic the COMPAS by **predicting the score-text** values with our model (The accuracy is around **75%**)
- We tried to **adjust the intercept and coefficient values** of all the features used in the logistic regression (as suggested by Propublica)
- We found that **Black defendants are 44% more likely** than white defendants to receive a higher score correcting for the seriousness of their crime, previous arrests, and future criminal behavior
- We found that the False Positive Rate **(FPR) for a black defendant is almost 90% higher** than a white defendant. It means a black defendant is 90% more likely to be falsely classified as high-risk criminal even though has not been rearrested in 2 years.

APPROACH 1 : FOR FAIRNESS

We decided to remove the features like Age category, Race and Sex from the features.

Instead used only charge-degree and priors-count.

The accuracy of the model decreased from 75 to 65 (since we are using less features)

The number of false positives has decreased for both the black and white BUT the FPR for the black remains almost 2 twice of that of white.

Important to mention : We consider that an **equal FPR** for both black and white means **fairness**

So omitting uncontrollable and sensitive data like race,sex and Age-category does not necessarily bring fairness.

APPROACH 2

We will try to optimize at training time.

We consulted the scientific paper **[Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment,Zafar et.al.(2016), page 6]**

We will try to implement something similar to **Covariance vs FPR** where decreasing the covariance threshold converges the FPR for both groups.

Here instead of covariance we will use a linear tuning parameter which on decreasing converges the FPR.

The main idea

We are forcing our classifier to mark high-risk to those black candidates only if the predicted probability of obtaining high-risk (calculated using predict_proba of sklearn) is greater that a certain threshold.

And this certain threshold is known as black threshold which is higher.

Subsequently if black_threshold is high then others_threshold will be lower because we defend black+other = 1.

This means that non black people will be marked as high-risk by our classifier for small values of predicted probability.

RESULTS

```
STATISTICS FOR WHITE CRIMINALS
----------------------------------------------
guessed  False  True
actual
False     1143   138
True       583   239
The False Positive Rate 0.10772833723653395
The Accuracy is 0.6571564431764146
STATISTICS FOR BLACK CRIMINALS
----------------------------------------------
guessed  False  True
actual
False     1182   332
True       871   790
The False Positive Rate 0.2192866578599736
The Accuracy is 0.6211023622047244
```

```
] Tuning_parameter = 0.45
  FairnessEnforcer(Tuning_parameter,1-Tuning_parameter,lr)
```

```
STATISTICS FOR WHITE CRIMINALS
------------------------------------------------------------------------------
guessed  False  True
actual
False     1078   203
True       516   306
The False Positive Rate 0.15846994535519127
The Accuracy is 0.6581074655254399
STATISTICS FOR BLACK CRIMINALS
------------------------------------------------------------------------------
guessed  False  True
actual
False     1255   259
True       986   675
The False Positive Rate 0.17107001321003962
The Accuracy is 0.6078740157480315
```

SUMMARY OF RESULTS

For the simp.classifiers	FPR of Black = Almost 2*FPR of White	Accuracy of the model: 65%
For logistic regression (inc. Race)	FPR of Black = Almost 4*FPR of White	Accuracy of the model: 75%
For logistic regression (Ex r,a,g)	FPR of Black = Almost 2*FPR of white	Accuracy of the model 64%
For logistic regression (our classifier)	FPR of Black = Almost 1.08*FPR of white	

CONCLUSION

So now our classifier has :

- Low accuracy for black criminals (Limitation : Accuracy less = tradeoff)
- Low number of False positives for black criminals (FNR of Black criminals increasing : Not good)
- FPR of Black = 1.08 * FPR of White
- FPR of white is increasing (Isn't this injustice for the white???)

To achieve fairness = optimise multiple fairness notions simultaneously. We optimized only 1

FAIRNESS IN A CLASSIFIER SOUNDS EASY BUT IT IS VERY DIFFICULT TO IMPLEMENT :)