Who is ready to leave

-A study of judgement of Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)

By: Lingxuan Shi, Chuanbo Tang

Introduction:

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.
It takes several features as input and give a score for each criminal how likely they will reoffend. Also, the score is labeled high, medium, and low of recidivism. And then the court will assign according to judgement to each defendant.

Lately, the fairness of COMPAS has been questioned. For example, VERNON PRATER, a Caucasian male who offended 2 armed robberies, 1 attempted armed robbery and 1 grand theft, was scored 3 and low risk; while BRISHA BORDEN, an African American female who offended 4 juvenile and misdemeanors was score 8 and high risk. Two years later, Borden has not been charged with any new crimes. Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars' worth of electronics.
This situation happens frequently across the nation, so we want to test the tendentiousness and fairness of COMPAS and see whether our typical Machine learning model could do a better job in this field.

Questions that we want to answer:

- Is there any Racial Bias in Compas (Risk of Recidivism and Risk of Violent Recidivism)?
- Is Compas fair for all races?
- Can we fit a typical Machine Learning model to predict the Risk of Recidivism?

Background reference:

Fairness

https://www.cs.toronto.edu/~toni/Papers/icml-final.pdf

Bias of COMPAS

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Logistic regression

https://www.econstor.eu/bitstream/10419/86100/1/02119.pdf

Ensemble (bagging, boosting)

https://arxiv.org/abs/2104.02395

Dataset:

We get the data from the GitHub page of compas-analysis project (https://github.com/propublica/compas-analysis/)

And their datasets are from Broward County because it is a large jurisdiction using the COMPAS tool in pretrial release decisions and Florida has strong open-records laws.

Because Broward County primarily uses the score to determine whether to release or detain a defendant before his or her trial, they discarded scores that were assessed at parole, probation or other stages in the criminal justice system. 11,757 people who were assessed at the pretrial stage are left.

They matched the criminal records to the COMPAS records using a person's first and last names and date of birth. This is the same technique used in the Broward County COMPAS validation study conducted by researchers at Florida State University in 2010. Then they downloaded around 80,000 criminal records from the Broward County Clerk's Office website.

To determine race, they used the race classifications used by the Broward County Sheriff's Office, which identifies defendants as black, white, Hispanic, Asian, and Native American. In 343 cases, the race was marked as Other.

They also compiled each person's record of incarceration. Jail records from the Broward County Sheriff's Office from January 2013 to April 2016, and they downloaded public incarceration records from the Florida Department of Corrections website.

After merging all tables together, we get a dataset with recidivate as outcome and race, gender, prior offense count, jail time, age etc. as X variables.

Method:

To answer the first question, we will use d Logistics regression to git a model on X variables and Decile score to find the odds ratio of the probability that African Americans are scored as high risk compared to Caucasians for both Risk of Recidivism and Risk of Violent Recidivism.

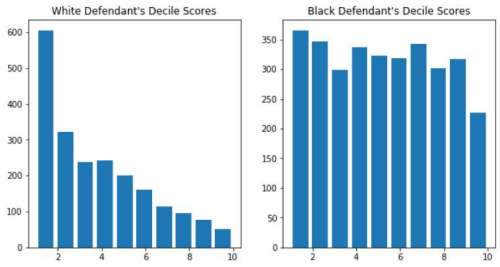
Secondly, we will do fairness test on accuracy equality, predictive parity, calibration, conditional use accuracy equality (equal PPV and NPV) and conditional procedure accuracy equality (equal FPR and FNR).

Then we will apply several ML algorithms of classification to predict the risk of each prisoner to see if we could get a better result than COMPAS. We will include algorithms such as bagging, boosting and random forest to avoid overfitting.

Data preprocessing and statistical analysis:

The raw data contains columns like name and offense date that are obviously not related to the outcome, so we firstly pick out those columns that might be statistically significant, including gender, race, age, age categories (under 25, 25-40, above 40), whether have recidivated before or not, jail time etc. Total number of rows is 7214.

Then we look at the data carefully. There are 809 rows with null in 4743 rows. After picking out those rows with NaN, we found that most NaN is from score_text col. And we also want to remove: rows whose charge date of a defendants Compas scored crime was not within 30 days from when the person was arrested; rows with re_recid = -1, which means no compas case at all; rows of ordinary traffic offenses -- those with a c_charge_degree of 'O' -- will not result in Jail time. There are 6172 rows left.



From above graph we could see that AA defendants have a more uniform distribution of decile scores; while white defendants have more low decile scores.

We then fit a logistics regression model and pick out those significant variables.

```
control = np.exp(-1.3241) / (1 + np.exp(-1.3241))
np.exp(-1.0834) / (1 - control + (control * np.exp(-1.0834)))
```

```
0.3930891806836198
```

Logit Regression Results

Dep. Variable:	y_bias	**No. Observations:**	3821
Model:	Logit	**Df Residuals:**	3814
Method:	MLE	**Df Model:**	6
Date:	Mon, 01 Nov 2021	**Pseudo R-squ.:**	0.2212
Time:	16:14:09	**Log-Likelihood:**	-1763.2
converged:	True	**LL-Null:**	-2263.9
Covariance Type:	nonrobust	**LLR p-value:**	4.548e-213

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.3241	0.125	-10.616	0.000	-1.569	-1.080
race[T.Caucasian]	-1.0834	0.092	-11.730	0.000	-1.264	-0.902
sex[T.Male]	0.1348	0.114	1.178	0.239	-0.089	0.359
age_cat[T.Greater than 45]	-1.0813	0.133	-8.140	0.000	-1.342	-0.821
age_cat[T.Less than 25]	0.5824	0.098	5.943	0.000	0.390	0.774
is_recid[T.1]	1.6158	0.088	18.323	0.000	1.443	1.789
c_charge_degree[T.M]	-0.7448	0.093	-8.013	0.000	-0.927	-0.563

We can see that white defendants are 60% less likely than Black defendants to receive a higher score correcting for the seriousness of their crime, previous arrests, and future criminal behavior.

And then we do the same analysis to the Risk of Violent Recidivism dataset.

White defendants are 63% less likely than Black defendants to receive a higher violent score correcting for the seriousness of their crime, previous arrests, and future criminal behavior.

Dep. Variable:	recy_bias	**No. Observations:**	2547
Model:	Logit	**Df Residuals:**	2540
Method:	MLE	**Df Model:**	6
Date:	Fri, 29 Oct 2021	**Pseudo R-squ.:**	0.2221
Time:	19:04:01	**Log-Likelihood:**	-971.26
converged:	True	**LL-Null:**	-1248.6
Covariance Type:	nonrobust	**LLR p-value:**	1.384e-116

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.4767	0.157	-9.391	0.000	-1.785	-1.168
race[T.Caucasian]	-1.1148	0.129	-8.640	0.000	-1.368	-0.862
sex[T.Male]	0.2106	0.154	1.367	0.172	-0.091	0.512
age_cat[T.Greater than 45]	-1.1958	0.193	-6.211	0.000	-1.573	-0.818
age_cat[T.Less than 25]	0.8419	0.134	6.283	0.000	0.579	1.105
is_recid[T.1]	1.7014	0.122	13.999	0.000	1.463	1.940
c_charge_degree[T.M]	-0.5770	0.125	-4.614	0.000	-0.822	-0.332

```
control = np.exp(-1.4767) / (1 + np.exp(-1.4767))
np.exp(-1.1148) / (1 - control + (control * np.exp(-1.1148)))
```

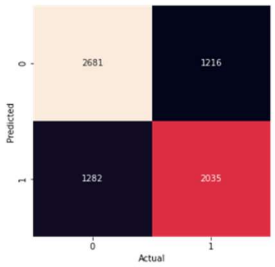
```
0.3748121838573907
```

Important Assumption:

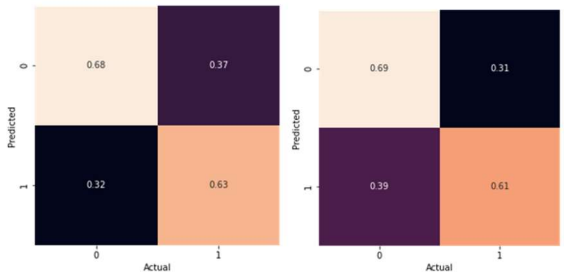
There is an assumption throughout our analysis, the predicted value is a categorical scores **1** to **10**, and we choose threshold to be **5** so scores below 5 means the criminal will not reoffend and scored large and equal to **5** will reoffend.

Confusion Matrix:

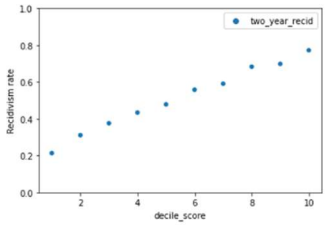
The picture below is the raw confusion matrix of the actual reoffend record and predicted value. 2681 and 2035 are the TP and TN, 1216 and 1282 are the FP and FN.

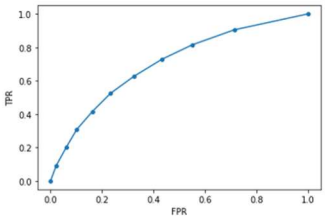


These 2 pictures are the confusion matrix after normalization. The left on is normalized for rows and the matrix on the right is normalized by columns. As we can see, the 2 matrix is close. Overall, we see that a defendant has a similar likelihood of being wrongly labeled a likely recidivist and of being wrongly labeled as unlikely to re-offend:



Then we have this plot shows the relation of decile scores and true recidivism rate. Defendant with higher compass score indeed have higher rates of recidivism.





This is the AUC score of the ROC curve for COMPAS algorithm. It is around 0.7 which is acceptable.

Fairness test:

Race is not an explicit input to COMPAS, but some of the questions that are used as input may have strong correlations with race.

Overall accuracy equality

As we can see from picture below, the TP + TN rates of COMPAS prediction for both races are quite similar. Thus, the overall accuracy equality is satisfied.

```
race
African-American    0.638258
Caucasian           0.669927
dtype: float64
```

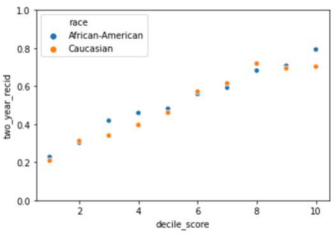
Predictive parity

This property is reflected by computing positive predictive value just like what we have learnt in class. Again, the difference is still within a few points, and this predictive parity fairness is also satisfied.

```
race
African-American    0.629715
Caucasian           0.591335
Name: two_year_recid, dtype: float64
```

Calibration

Similar to predictive parity, we check whether a defendant with a given score has the same probability of recidivism for the two groups. As we can see from the plot below, African American and Caucasian perform quite similar for all given decile scores, recidivism rate is similar. And this is the fairness called calibration.



Existing controversy

However, according to what we observed in this dataset, white defendant tends to have lower decile scores than black defendants. We can see from the picture that there are much more white defendants labeled score1 than black defendants and more black defendants marked score 10 than white defendants. To test if this bias, we need to check the real prevalence in the dataset, whether the rate of recidivism is the same for both groups.

Now we see the real reoffend rate for AA is 0.51, cacasian is 0.39, and the predicted rate is 0.58 and 0.34 for AA and caucation. The predictied value is not that close to the rate in our previous analysis, but it is still close.

race	two_year_recid	is_med_or_high_risk	decile_score
African-American	0.514340	0.588203	5.368777
Caucasian	0.393643	0.348003	3.735126

So far, according to our analysis based on overall accuracy equality, prediction parity, and calibration, COMPAS seems to be quite fair.

But the predicted score is a bit higher for AA and lower for CA.

We find a way to fix this, that is changing the threshold for each group that we mark as will reoffend or will not reoffend.

```
False positive rate (white)      :  0.350134408602150 56     False positive rate (Black)      :  0.34317548746518106
False negative rate (white)      :  0.360248447204968 96     False negative rate (Black)      :  0.37243556023145713
Positive predictive value (white):  0.5425812115891132      Positive predictive value (Black):  0.6594803758982863
Negative predictive value (white):  0.7353612167300381      Negative predictive value (Black):  0.6248012718600954
```

Now we change the threshold for AA down to 4 and CA up to 6, now we have FP FN rate similar for both group, which is the equalized odds fairness.

We have worked so hard trying to satisfy all fairness property at once, however, due the conclusion we have learnt in class, impossibility theorem for fairness metrics. There is no method that can satisfy all fairness.

Apply ML algorithm to predict recidivism:

We apply each ML algorithm on predicting recidivism (thus classification). Our X variables are age, sex, recidivate or not, charge degree, priors jail count, days before screening arrest and jail time.

For numeric variables, we normalize them to make sure they have the effect on measure of distance.

For categorical variables, we apply One Hot encoding to turn them into numerical variables.

We also apply 10-fold cross validation to avoid overfitting or selection bias. And for all algorithms, we use for loops to find out the best parameters (n_estimator/K/ penalty) that give us the best accuracies.

For DecisionTree and KNN, we use bagging to reduce the probability of overfitting even further. The results are below.

Algorithm	**Accuracy**	**Best n_estimator/K/ penalty**
Bagging DecisionTree	0.910	10
Random Forest	0.927	9
Adaboost	0.933	9
Bagging KNN	0.920	6
Logistic Regression	0.934	L2

From the above table we could see that all ML algorithms have decent accuracy, while Adaboost and Logistics regression are the best two with accuracy over 93%.

It may imply that we could rely more on the judgement of the machine instead of giving an obscure score to the court.

Conclusion:

In our analysis, COMPAS algorithm is fair in many metrics but fail in statistical parity within subgroups. This might be caused by the base rate difference in these 2 groups reflected by a single algorithm. We cannot say there won't be another algorithm that can cover most metric including statistical parity, but we still cannot find a relative better one than COMPAS.

Such base rate differences can cascade through fairness test and lead to difficult tradeoffs.

On the other hand, our ML models do a great job in predicting recidivism, but this method may need more ethic test and discussion compared to judgement that made by the court.