Data Science Project - COMPAS

MA Yuqiang & TUNG Phan Manh & WU Zhongyuan

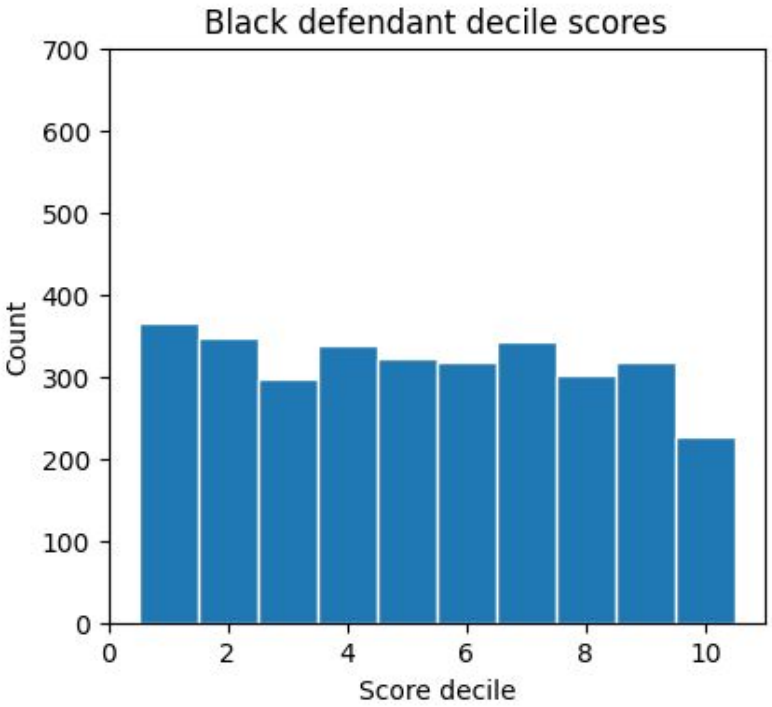
Introduction

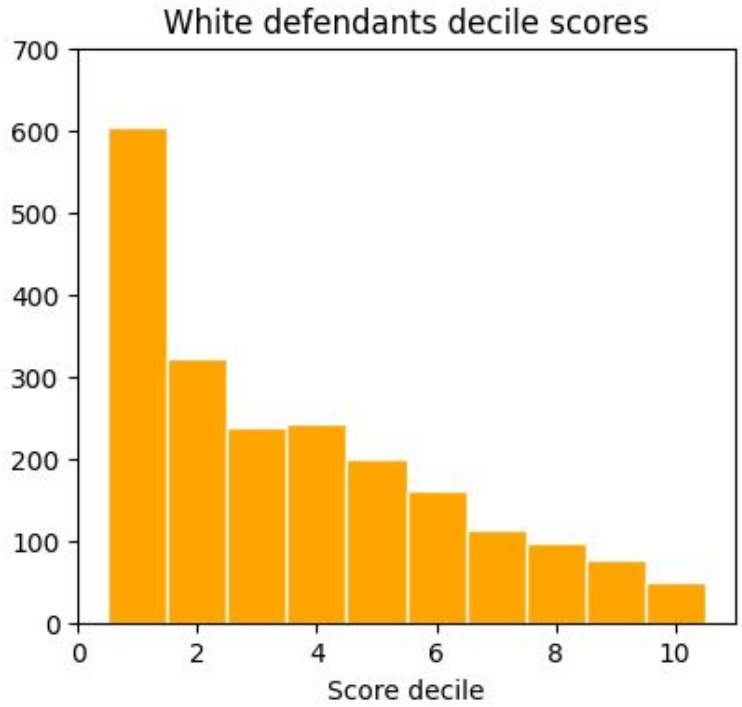
- **ProPublica COMPAS dataset.**
- Controversial dataset consisting of over 10,000 criminal defendant profiles in Broward County, Florida, USA.
- Various features of the defendant: criminal history, demographics, and COMPAS scores.
- COMPAS is the tool used to assess the likelihood of a defendant committing another crime in the future.
- However, the COMPAS scores have been shown to have biases against certain racial groups.
- Therefore, we need a “Fair” classifier which can decrease these biases.

Focused attributes in COMPAS dataset

- **age**: The age of the individual at the time of their arrest or booking.
- **c_charge_degree**: The degree of the criminal charge, either felony or misdemeanor.
- **race**: The race of the individual.
- **sex**: The gender of the individual.
- **priors_count**: The number of prior criminal convictions the individual has.
- **two_year_recid**: A binary variable indicating whether the individual was re-arrested within two years of their initial arrest.
- **score_text**: A text-based score from the COMPAS algorithm indicating the individual's risk of recidivism.
- **decile_score**: A numerical score from the COMPAS algorithm indicating the individual's risk of recidivism, on a scale from 1 to 10.

Exploration





Pipeline: Normal & Modified

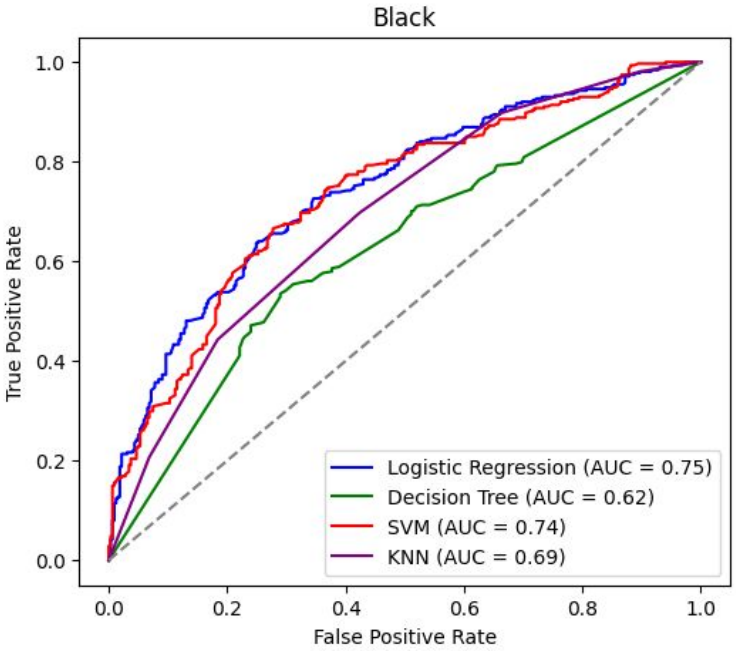
- Input: age, race, sex, priors_count, c_charge_degree
- Output: two_year_recid
- Pre-processing: convert categorical data, split train-test 70-30
- Modelling: Logistic Regression
- Post-processing: Receiver Operating Characteristic post-processing (p=0.4)

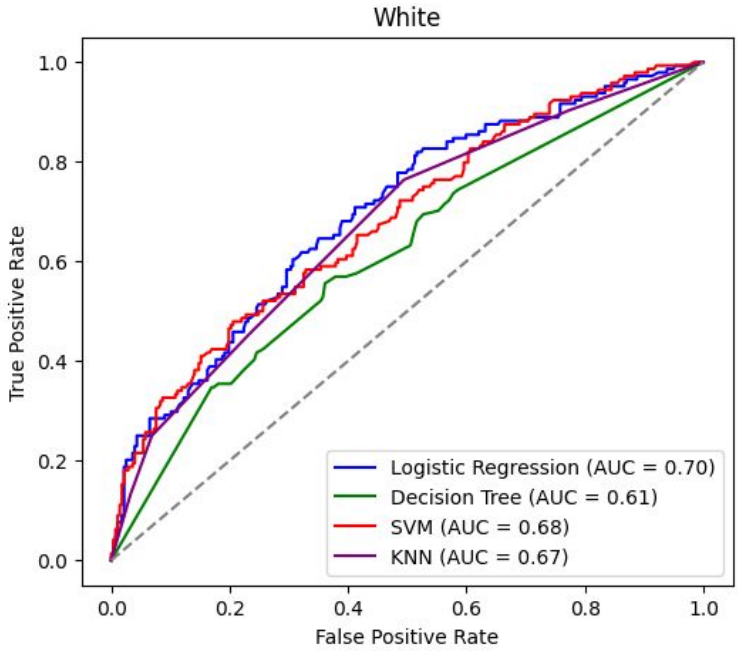
```
def roc_postprocessing(y_pred_probs, p):
    # Compute the optimal threshold for the ROC post-processing algorithm
    n = len(y_pred_probs)

    # Find optimum threshold for each group; higher p more fairness lower accuracy
    threshold = np.percentile(y_pred_probs, (1 - p) * 100)

    # Apply the threshold to obtain binary predictions
    y_pred = (y_pred_probs > threshold).astype(int)
    return y_pred
```

Fairness Metrics

- 3 core laws (metrics) for the Fairness:
  - Independence:
    i. Equal proportion of positive outcomes ($\hat{Y} = 1$) in each population.
    ii. The lower difference of the value (TP + FP)/(TP + FP + TN + FN) among the ethnic groups, the better.
  - Separation:
    i. Equal false positive/negative rates ($\hat{Y} \not\perp Y$) in each population.
    ii. The lower difference if the FPR and FNR among the ethnic groups, the better.
  - Sufficiency:
    i. Equal true positive/negative rates in each population.
    ii. The lower difference if the TPR and TNR among the ethnic groups, the better.
- Core function: *evaluate_fairness()*
- It exists contradictions among the laws, we can satisfy at most 2 laws at the same time.

Model performance comparison





Model fairness comparison

LogisticRegression:

```
---------Independence performance---------
positive rate 1 is 0.228029, positive rate 2 is 0.532283, difference is 0.304255
---------Separation performance---------
FPR 1 is 0.166065, FPR 2 is 0.364486, difference is 0.198421
FNR 1 is 0.652778, FNR 2 is 0.296178, difference is 0.356599
---------Sufficiency performance---------
TPR 1 is 0.347222, TPR 2 is 0.703822, difference is 0.356599
TNR 1 is 0.833935, TNR 2 is 0.635514, difference is 0.198421
```

Decision Tree:

```
---------Independence performance---------
positive rate 1 is 0.251781, positive rate 2 is 0.485039, difference is 0.233258
---------Separation performance---------
FPR 1 is 0.184116, FPR 2 is 0.330218, difference is 0.146103
FNR 1 is 0.618056, FNR 2 is 0.356688, difference is 0.261368
---------Sufficiency performance---------
TPR 1 is 0.381944, TPR 2 is 0.643312, difference is 0.261368
TNR 1 is 0.815884, TNR 2 is 0.669782, difference is 0.146103
```

SVM:

```
---------Independence performance---------
positive rate 1 is 0.256532, positive rate 2 is 0.555906, difference is 0.299373
---------Separation performance---------
FPR 1 is 0.184116, FPR 2 is 0.376947, difference is 0.192832
FNR 1 is 0.604167, FNR 2 is 0.261146, difference is 0.343020
---------Sufficiency performance---------
TPR 1 is 0.395833, TPR 2 is 0.738854, difference is 0.343020
TNR 1 is 0.815884, TNR 2 is 0.623053, difference is 0.192832
```

KNN:

```
---------Independence performance---------
positive rate 1 is 0.315914, positive rate 2 is 0.566929, difference is 0.251015
---------Separation performance---------
FPR 1 is 0.245487, FPR 2 is 0.429907, difference is 0.184419
FNR 1 is 0.548611, FNR 2 is 0.292994, difference is 0.255617
---------Sufficiency performance---------
TPR 1 is 0.451389, TPR 2 is 0.707006, difference is 0.255617
TNR 1 is 0.754513, TNR 2 is 0.570093, difference is 0.184419
```

Model fairness comparison

	Independence diff	Separation (FPR) diff	Sufficiency (TPR) diff
Logistic Regression	0.304255	0.198421	0.356599
Decision Tree	0.233258	0.146103	0.261368
Support Vector Machine (SVM)	0.299373	0.192832	0.343020
K nearest neighbors (KNN)	0.251015	0.184419	0.255617

=> Decision Tree, best on the fairness!

Result of fair classifier

The result of a Normal Classifier

```
==========Normal Classifier Result==========
--------Result of Print_Metrics in group 1--------
Accuracy:  0.6687797147385103
PPV:  0.5828220858895705
FPR:  0.17215189873417722
FNR:  0.597457627118644

--------Result of Print_Metrics in group 2--------
Accuracy:  0.6505771248688352
PPV:  0.6687370600414079
FPR:  0.350109409190372
FNR:  0.3487903225806452

--------Independence performance--------
positive rate 1 is 0.258320, positive rate 2 is 0.506821, difference is 0.248500
--------Separation performance--------
FPR 1 is 0.172152, FPR 2 is 0.350109, difference is 0.177958
FNR 1 is 0.597458, FNR 2 is 0.348790, difference is 0.248667
--------Sufficiency performance--------
TPR 1 is 0.402542, TPR 2 is 0.651210, difference is 0.248667
TNR 1 is 0.827848, TNR 2 is 0.649891, difference is 0.177958
```

The result of a Fair Classifier

```
==========Modified Result==========
--------Result of Print_Metrics in group 1--------
Accuracy:  0.6545166402535658
PPV:  0.5357142857142857
FPR:  0.29620253164556964
FNR:  0.4279661016949153

--------Result of Print_Metrics in group 2--------
Accuracy:  0.6484784889821616
PPV:  0.7267605633802817
FPR:  0.212253829321663
FNR:  0.4798387096774194

--------Independence performance--------
positive rate 1 is 0.399366, positive rate 2 is 0.372508, difference is 0.026858
--------Separation performance--------
FPR 1 is 0.296203, FPR 2 is 0.212254, difference is 0.083949
FNR 1 is 0.427966, FNR 2 is 0.479839, difference is 0.051873
--------Sufficiency performance--------
TPR 1 is 0.572034, TPR 2 is 0.520161, difference is 0.051873
TNR 1 is 0.703797, TNR 2 is 0.787746, difference is 0.083949
```

Thank you for listening!
Q & A time