Algorithm Bias in COMPAS

Wiley Winters

MSDS 640 Ethics, Privacy, and Social Justice in Data Science

2025-Aug-24

Introduction

This case study examines, analyzes, and models the racial bias in the Correctional Offenders Management Profiling for Alternative Sanctions (COMPAS) algorithm.

A 2016 ProPublica report highlighted that COMPAS has a pattern of assigning higher risk scores to African American offenders when compared to Caucasians and other racial groups

My aim is to confirm or refute ProPublica’s findings.

Research Question

This study will use the ProPublica datasets to validate if there is bias in COMPAS algorithms or not.

Significance of Study

REGIS UNIVERSITY

Using biased algorithms in judicial decision-making brings up a few ethical concerns:

- Unjust incarceration of individuals based on membership in certain groups
- Erosion of trust in the legal system
- Lack of transparency; many algorithms are proprietary and considered trade secrets

Methodology

For this study a pronged approach will be used.

- Conduct a statistical analysis of the ProPublica dataset to prove or disprove Propublica’s findings
- Select and train a Machine Learning (ML) model using the ProPublica dataset to determine if such an approach is feasible and more accurate than the COMPAS algorithm when predicting two-year recidivism.

Throughout this study, items and conclusions from the review of literature will be references when appropriate

Data Set

- Dataset used is from the ProPublica repository. I selected the set with a two-year recidivism flag
- 7,214 rows and 53 features constituted the dataset
- Standard practices were followed when cleaning the data and missing values were imputed or dropped, datatype mismatches were fixed, and risk scores of -1 were removed
- After cleaning, the number of features were reduced to 28

Exploratory Data Analysis (EDA)

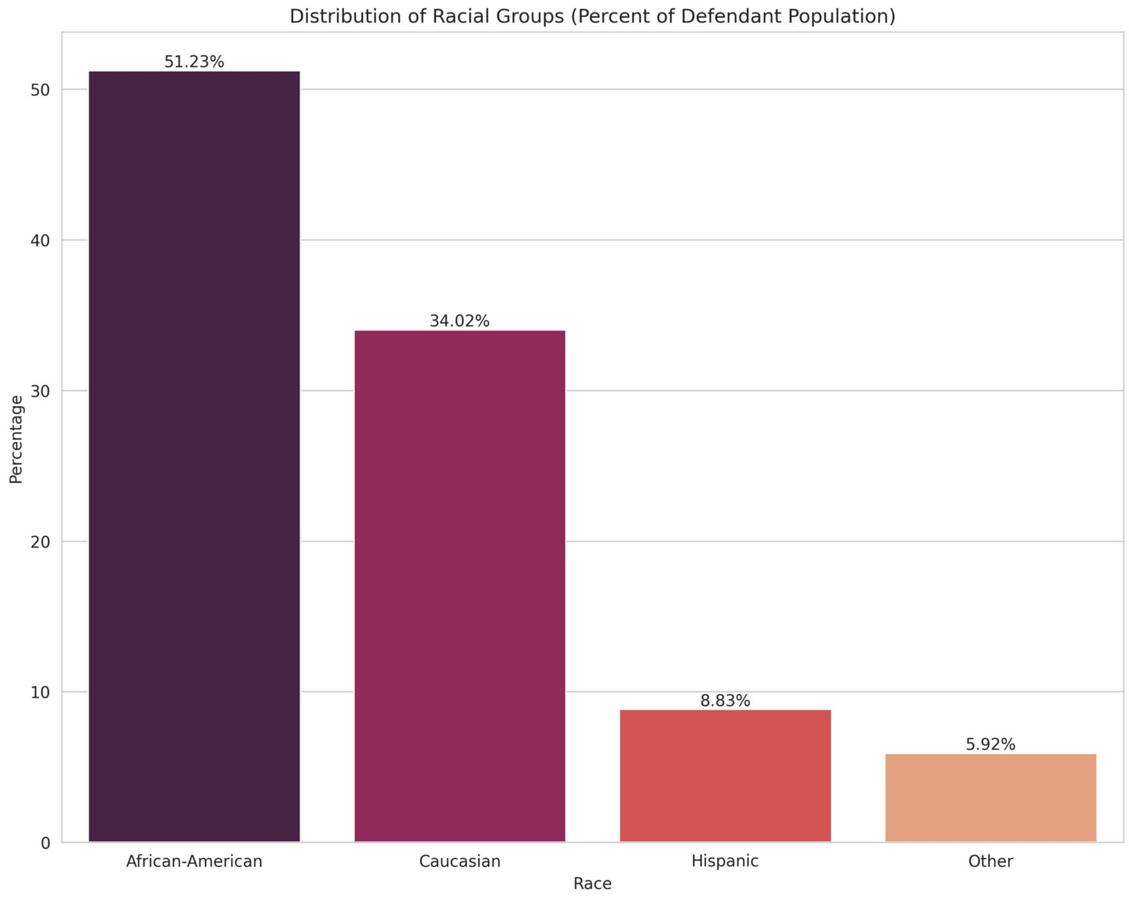
- EDA was conducted to identify any trends or patterns within the data
- Classes were mostly balanced with the exception of racial groups
  - African Americans over represented
  - Asian and Native Americans under represented
- Asian and Native Americans were combined into the *Other* group

Racial Group	**Number of Individuals**
African American	3,696
Caucasian	2,454
Hispanic	637
Other	427

Racial Group Distributions

From the start some concerning patterns emerged

- African Americans make up a little over 51% of the offender population followed by Caucasians
- In 2016 African Americans were only about 29% of the population where the dataset was obtained from
- This disparity indicates that bias begins before the individual enters the judicial system.

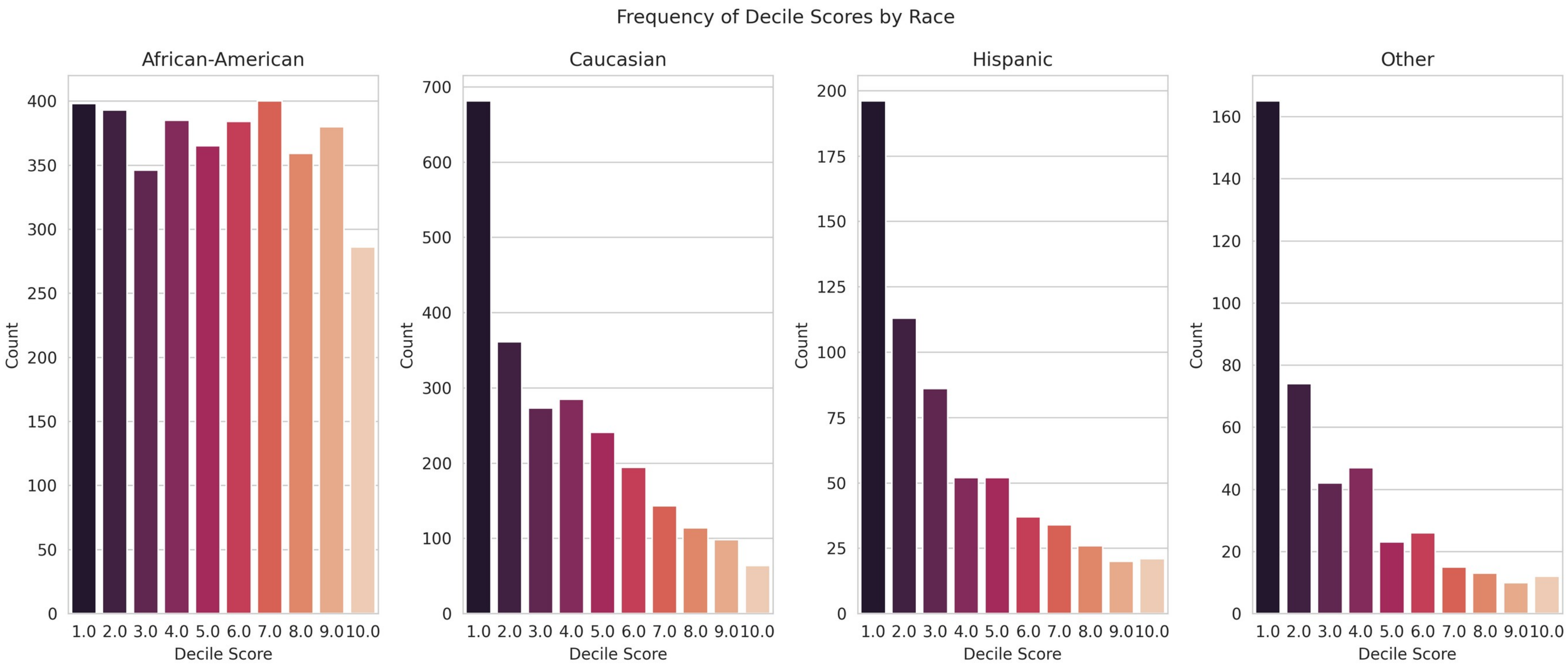


Analysis

- In the COMPAS system, the risk score assigned to a defendant is referred to as the *decile score*, which ranges from 1 to 10.
- A lower decile score corresponds to a lower risk.
- The ProPublica report highlighted that African Americans receive higher decile scores than other racial groups.
- My analysis of these scores support this conclusion

Racial Group	Mean	Standard Deviation
African American	5.37	2.83
Caucasian	3.74	2.60
Hispanic	3.46	2.60
Other	3.08	2.48

Analysis – Visualization of Decile Score Distribution



A visual representation of the decile score makes it easier to understand and compare the racial groups. As previously discussed, African Americans' decile scores are higher than the other groups.

ANOVA (Analysis of Variance)

REGIS UNIVERSITY

- To determine whether the differences in decile scores between African Americans and other groups are statistically significant; I performed a one-way ANOVA test.
- The decile scores were normalized using the *Min-Max* method to ensure all scores were on the same scale
- The *P-Value* is below *0.05* which indicates it is **statistically unlikely that the observed differences in mean decile scores between African Americans and other groups are due to chance alone**

```
# Normalize decile_score using Min-Max method
compas_df['normalized'] = compas_df.groupby('race')['decile_score']. \
          transform(lambda x: (x - x.min()) / (x.max() - x.min()))

# group by race on the normalized value
groups = compas_df.groupby('race')['normalized'].apply(list)

# Perform the oneway ANOVA test
f_stat, p_val = stats.f_oneway(*groups)

# Print results
print('F-statistic: ',f_stat)
print('P-value:     ',p_val)
if p_val < 0.05:
    print('Null hypothesis is rejected')
```

```
F-statistic:  261.01168051564656
P-value:      8.089670291495388e-161
Null hypothesis is rejected
```

Modeling

- A recent study suggested the algorithm used by COMPAS may have an accuracy rate as low as 68% when predicting the likelihood of recidivism
- To investigate this further, I used the COMPAS dataset to train three models with the two-year recidivism flag as the target class.
- The models trained include:
  - Logistic Regression
  - Support Vector Machines (SVM)
  - Random Forest Classifier
- Models were selected because:
  - They are commonly used in binary classification problems
  - Easy to implement
  - Easy to interpret

Modeling – Training Results

- All models performed well with this dataset
- The Random Forest Classifier over fitted the data, but actions were taken to compensate for this
- Since all models being evaluated performed about the same, a K-Fold cross-validation test was conducted using Negative Mean Square Error (NMSE) as the evaluation metric

Model	Train	Accuracy	TPR	FNR
Logistic Regression	0.88	0.89	0.92	0.56
Support Vector Machine	0.90	0.89	0.96	0.54
Random Forest Classifer	0.91	0.89	0.99	0.52

Modeling – K-Fold Results

- When using NMSE as the scoring metric, the model that scores closest to zero is typically considered the best performer.
- In this case the Random Forest Classifier's score was closest to zero

Model	NMSE	Standard Deviation
Logistic Regression	-0.296990	0.121847
Support Vector Machine	-0.296833	0.014517
Random Forest Classifier	-0.101370	0.007385

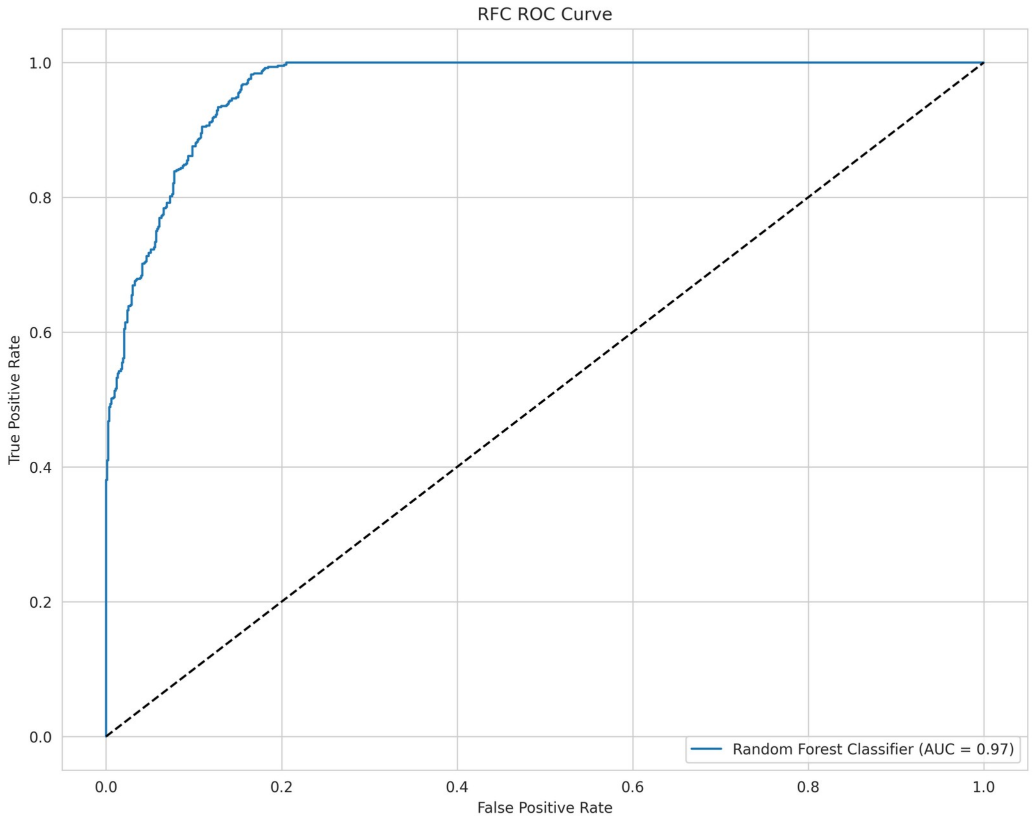
Results

REGIS UNIVERSITY

- This analysis found that recidivism is higher within the African American population in this study; however, this group also made up over half of the defendant population
- African Americans were slightly above 51% of the defendant population, but only comprised a little over 28% of the total population of the county where the COMPAS data was obtained
- Caucasians represented only about 34% of the defendant population, but comprised nearly 61% of the county's population
- It is essential to recognize that the disparity observed in this study begins well before individuals enter the judicial system.

Results – Modeling

- Based on its NMSE score, the Random Forest Classifier was selected for further evaluation and testing.
- AUC (Area Under Curve) is the preferred method of evaluating the performance of a binary classifier model.
- The Random Forest Classifier with some minor hyperparameter tuning achieved an AUC 0.97



Random Forest Classifier Performance

- The Random Forest Classifier was slightly over fitted and required hyperparameter tuning to compensate
- A training score of 1.00 (100%) is the indicator of this.
- After applying the adjusted hyperparameters, the training score become more acceptable and TPR improved while FNR decreased

Metric	Before	After
Train Score	1.00	0.91
Accuracy Score	0.91	0.89
TPR	0.95	$\boxed{0.99}$
FNR	0.55	$\boxed{0.52}$

Discussion and Conclusions

- My analysis of ProPublica's COMPAS dataset confirms the pattern identified in ProPublica's study and COMPAS does have a bias toward awarding higher decile scores to African Americans
- Other studies have also found bias in the algorithm, but offer different interpretations on it.
- One study suggested that COMPAS reduces gender bias since judges tend to be more lenient with female offenders, whereas COMPAS is not
- The model creating and evaluation I performed indicates a commonly used free machine learning model can be trained on recidivism data and produce results comparable or exceeding that of the original COMPAS algorithm.

Questions