2025

Bias in Predictive Modelling for Public Safety: Evaluating Fairness Across Protected Characteristics in Crime Data

E4216209
AJAKAYE, JESUBUKADE

BODY WORD COUNT: 2055

Bias in Predictive Modelling for Public Safety: Evaluating Fairness Across Protected Characteristics in Crime Data

By

AJAKAYE, JESUBUKADE
School of Computing, Engineering and Digital Technologies,
Teesside University

Abstract

Ethical concerns have been a recurring theme for intelligent models that would be applied to human society. Some of these models inherit bias innate in the society eroded trust instead of acceptance. In predictive modelling, biases can be transmitted through out the machine learning pipeline from data collection to model design to deployment. This study aimed to understand and address bias in predictive modelling in security and public safety using selected protected characteristics. A Random Forest Model was developed to predict the statistical murder flag of victims in a shooting crime record from New York City. The model after optimisation performed with an accuracy of 68% before three (3) fairness criteria including equal accuracy, group fairness, and equality of opportunity were used to evaluate the model for bias. Finding revealed a bias in the model between Race (White and Black) and Sex (Man and Woman) While Age (Young and Adult) exhibited little or no bias. The study underscores the importance of ensuring ethical approach during and after model development and ensure the system is fair to all groups before eventual application to the society.

Keyword: AI Ethics, Bias, Fairness Criteria

1.0 Introduction

The safety of life and property is an important and a fundamental human right and as a result stakeholders saddled with the mandate to protect must ensure necessary action to mitigate crime or provision of rapid response to victims should crime occur. Various intelligent systems have been developed to predict crime in the criminal justice system or for rapid response but some of these systems are riddled with bias within various protected characteristic groups which as led to several debates and opinion on the application of these tools. Bias in predictive modelling refers to systematic errors that lead to unfair or inaccurate outcomes for certain groups, often due to issues in data collection, model design, or deployment (Montana, *et al*. 2023). Various studies have established bias in the attention and support given to crime victims based on their race, gender, age group among others. HMICFRS (2024) reported that provision of general support services to victims in the criminal justice system exhibit significant differences between areas in England and Wales regarding how long victims must wait to speak to someone. Consequently, if there is bias in tools used in criminal justice systems for crime detection and prediction, there is a high possibility that these biases will be evident in police response to victims and crimes based on these biases. Tools like PredPol, COMPAS among others have been found to be biased within various groups (Dressel and Farid 2018; Khademi and Honavar 2019; Mehrotra and Cameron 2023). These differences is treatment have resulted in avoidable deaths, physical and mental well-being of victims, and eroded trust in justice system and societal equity (Hanson, *et al*. 2010. Understanding and addressing bias in predictive modelling is crucial to developing fair and reliable AI systems therefore this research therefore investigates availability of bias in selected protected characteristics in the prediction of murder in shooting crime record.

2.0 Methodology

This section reports the approach the research took for the development of model and application of fairness criteria to check for bias in the model in relation to the victims of the crime. The phases of this methodology include, Data exploration and preprocessing, model

development, model optimisation and evaluation, and application of 3 fairness criteria.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 21626 entries, 74146165 to 193694862
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   OCCUR_DATE              21626 non-null  object
 1   OCCUR_TIME              21626 non-null  object
 2   BORO                    21626 non-null  object
 3   PRECINCT                21626 non-null  object
 4   JURISDICTION_CODE       21624 non-null  float64
 5   LOCATION_DESC           9265 non-null   object
 6   STATISTICAL_MURDER_FLAG 21626 non-null  bool
 7   PERP_AGE_GROUP          14321 non-null  object
 8   PERP_SEX                14355 non-null  object
 9   PERP_RACE               14355 non-null  object
 10  VIC_AGE_GROUP           21626 non-null  object
 11  VIC_SEX                 21626 non-null  object
 12  VIC_RACE                21626 non-null  object
 13  X_COORD_CD              21626 non-null  object
 14  Y_COORD_CD              21626 non-null  object
 15  Latitude                21626 non-null  float64
 16  Longitude               21626 non-null  float64
 17  Lon_Lat                 21626 non-null  object
dtypes: bool(1), float64(3), object(14)
memory usage: 3.0+ MB
```

Figure 1: Information of Columns in Dataset

The researched used the New York City Shooting Incident Dataset publicly available on Kaggle. The dataset contained approximately 21,600 (See Figure 1). The study evaluated 3 protected characteristics which are Race, Sex and Age Group.

2.1 Data Exploration and Preprocessing

Feature Selection: Feature selection helps to identify predictors. For the purpose of this study the following columns were dropped: LOCATION_DESC which is the description of location had a lot of missing values and contains values that are too generic. PERP_AGE_GROUP, PERP_SEX and PERP_RACE were removed to avoid leakage into the model. JURISDICTION_CODE dropped as values do not seem to carry any meaning. X_COORD_CD, Y_COORD_CD, Latitude, Longitude and Lon_Lat were all dropped as BORO and PRECINCT carries the same information as correlates.

INCIDENT_KEY	OCCUR_TIME	BORO	PRECINCT	STATISTICAL_MURDER_FLAG	VIC_AGE_GROUP	VIC_SEX	VIC_RACE	MONTH
74146165	EARLY MORNING	QUEENS	113	False	25-44	M	BLACK	08
66928846	NIGHT	BROOKLYN	67	True	45-64	M	BLACK	10
29114164	NIGHT	BROOKLYN	75	False	25-44	M	BLACK	05
85180336	AFTERNOON	BROOKLYN	81	False	25-44	M	BLACK	06
73405770	EARLY MORNING	BRONX	47	False	25-44	M	BLACK	06

Figure 2: Top Rows of Cleaned Dataset

Data Transformation: OCCUR_DATE was transformed into Month which the crime occurred while the OCCUR_TIME column was transformed into EARLY MORNING, MORNING, AFTERNOON or NIGHT to understand when shooting crime occurs most.

Figure 3 revealed that most shooting crime occur between 6pm and 6am and also recorded more fatality at this period.

Protected Characteristics: The study ensured that the protected characteristics to be considered were transformed into two groups to support model development and analysis.

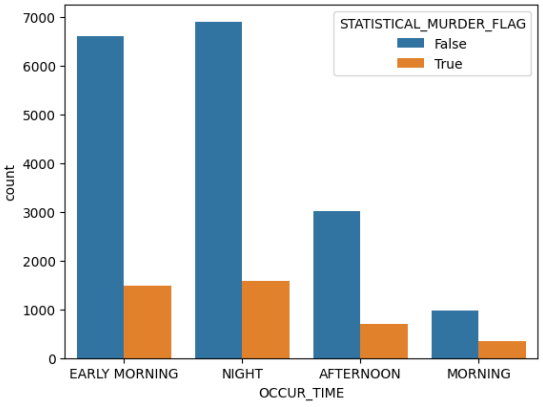

Figure 3: Distribution of Shooting Time by Murder Flag

Figure 4 revealed that the distribution spread of Black to White and Male to Female in the dataset is imbalanced while Adult to Young is relative balance. This showed that Blacks and Males are the most shooting victims.

Further analysis revealed that the proportion of Victims for Race and Sex seemed balanced while for Age Group, more Adult had more fatality when compared to Young victims (see Figure 5).

Figure 4: Distribution of Protected Characteristics based on Murder Flag

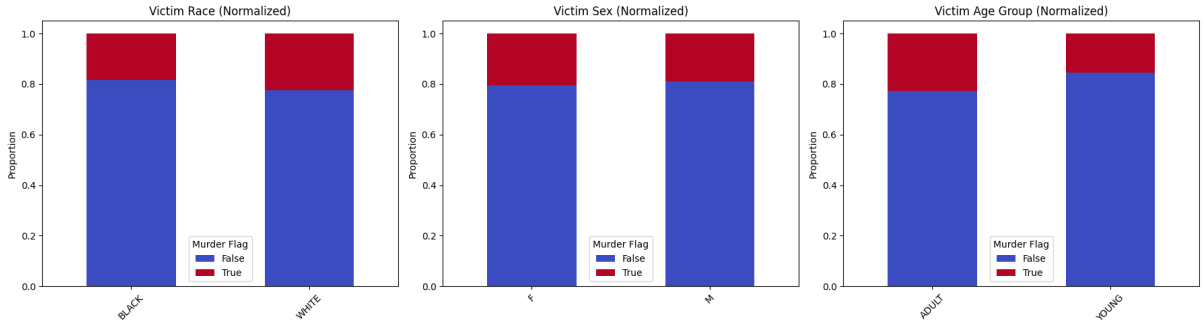

Figure 5: Proportion of Murder Flag in each Group

2.2 Model Development

The research performed a binary classification using Random Forest Classifier. Random Forest Classifier was selected because it is robust to overfitting handles categorical variables well with minimal preprocessing. Random Forest is easy to explain as a tree ensemble model to understand how the model arrived at its predictions. Label encoding was performed on all features as all predictors are categorical variables.

Data Balance: The need to balance the dataset arose as the researcher discovered that the model was not training and was predict only one class and achieved about 87% accuracy. This was because the data was imbalanced and model able to achieve high accuracy despite guess on one class. NearMiss package was used to undersample. NearMiss is a smart undersampling technique that selects majority class samples closest to the minority class, helping to preserve decision boundaries. It improves model focus on hard-to-classify instances without introducing synthetic data.

Train-Test Split: The research used 80% of the balanced dataset for training keeping 20% as Test set to evaluate the Model Performance.

2.3 Model Optimisation and Evaluation

The model was evaluated based on its accuracy, precision, and recall. **Accuracy** is the proportion of correctly predicted observations to the total observations, calculated as: $\frac{(TP+TN)}{(TP+TN+FP+FN)}$. **Precision**, also known as **Positive Predictive Value**, measures the proportion of true positive predictions among all positive predictions, calculated as: $\frac{TP}{(TP+FP)}$. **Recall**, also known as **Sensitivity** or **True Positive Rate**, measures the proportion of actual positive cases that were correctly identified, calculated as: $\frac{TP}{(TP+FN)}$. where TP is a true-positive value, FP is a false-positive value, TN is a true-negative value and FN is a false-negative value. The model was able to distinguish between both classes after the data was balanced returning a 65% accuracy, 70% on precision and 53% recall.

After hyperparameter tuning using StratifiedKFold and RandomizedSearchCV to optimize the model (where

best params are: 'n_estimators': 300, 'min_samples_split': 4, 'min_samples_leaf': 5), the model returned 68% accuracy, 70% precision and recall improved to 62%. Figure 6 shows the confusion matrix of the optimized model where 0 is the positive class FALSE and 1 is the negative class TRUE

Table 1: Performance of Model based on Selected Metrics

Metrics	Base Model	Optimized Model
Accuracy	65.35%	68.01%
Precision	70.20%	70.40%
Recall	53.34%	62.13%

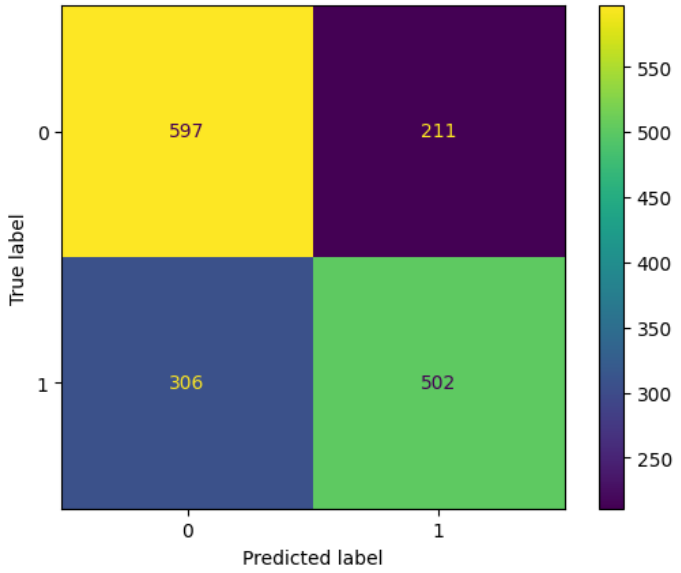

Figure 6: Confusion Matrix of Optimised Model

2.4 Fairness Criteria

To understand bias within the protected characteristics, the study evaluated the model predictions on:

- Equal Accuracy is applied to ensure the model performs equally well (in terms of overall accuracy) across different subgroups (e.g., race, gender, age). This suggest that the model may be significantly bias to the group with significantly lower accuracy. This is measured with each group accuracy $\frac{(TP + TN)}{(TP + TN + FP + FN)}$
  Equal accuracy alone doesn't reveal whether errors are systematically skewed (e.g., more false positives for one group).
- Group Fairness (demographic/statistical parity) ensures the concerned outcomes (e.g., predicted as "Possible Murder") are distributed equally across groups. Measured with each groups positive rate $\frac{(TP + FP)}{(TP + TN + FP + FN)}$
  The implication of statistical parity is that it can play down on merit to satisfy or justify that both groups are equally represented.
- Equality of Opportunity is a metric that ensure individuals who truly belong to the positive class are equally likely to be correctly identified across groups. Equality of Opportunity is measured with the Recall $\frac{TP}{(TP + FN)}$ of prediction in each group.
  Unequal recall suggests that some groups are systematically under-identified for positive outcomes, which may result in missed support or protection. The drawback of Equality of Opportunity is that it does not account for False Positives but only True Positive rates.

Indices of every group was used to locate their respective actual and predicted cases then confusion matrix was carried out for each group for the application of the fairness criteria. The 80% rule was adopted to check if there is bias within the group based on the fairness criteria applied. The 80% rule also called the four-fifth rule was developed to avoid discrimination in employment by the Equal Employment Opportunity Commission (EEOC) is a statistical reference if there is a substantial difference between two groups for any protected characteristics. The rule can also be applied to check for bias in machine learning models to check for substantial differences in the protected characteristics understudy.

Equal Accuracy

The overall model accuracy for the 3 protected characteristics showed no bias as represented on Table 2. The model was able to make approximately about the same correct classification for each group within the protected characteristics with Race having ratio 0.93, Sex having a ratio of 0.81 and Age Group having approximately ratio 1 correct prediction in both groups.

Table 2: Evaluating Accuracy across Groups

Protected Characteristic	Group	Accuracy	80% Rule Ratio	Comment
Race	Black	67.27%	92.90%	No bias identified
	White	72.41%		
Sex	Male	66.86%	80.97%	No bias identified
	Female	82.57%		
Age Group	Young	71.93%	99.81%	No bias identified
	Adult	71.79%		

Group Fairness

To ensure that the model is fair to both groups in each protected characteristics, the predicted positives was compared, that is, both groups received the same number of Murder alert irrespective. High disparity was seen in the attention given to the White population and Female populations in the model. This shows evidence of societal bias and how they are transferred to model development and training (see Table 3).

Table 3: Evaluating Positive Rate across Groups

Protected Characteristic	Group	Positive Rate	80% Rule Ratio	Comment
Race	Black	34.97%	35.43%	Bias against the Black
	White	98.71%		
Sex	Male	39.30%	39.30%	Bias against the Female
	Female	100%		
Age Group	Young	30.50%	74.31%	Bias against the Young
	Adult	41.04%		

Equality of Opportunity

Equality of opportunity focuses on equalizing true positive rates across different demographic groups, ensuring that qualified individuals have equal chances of receiving positive outcomes. Each data points are considered and ensures that individuals deserving gets the attention they deserve irrespective of the group they are in. This is evaluated by considering the proportion of predicted positives for each group against the actual positives for each group (see Table 4).

Table 4: Evaluating Recall across Groups

Protected Characteristic	Group	Recall	80% Rule Ratio	Comment
Race	Black	52.43%	53.36%	Bias against the Black
	White	98.25%		
Sex	Male	56.46%	56.46%	Bias against the Female
	Female	100%		
Age Group	Young	52.62%	84.69%	No bias identified

Confusion Matrix of all groups

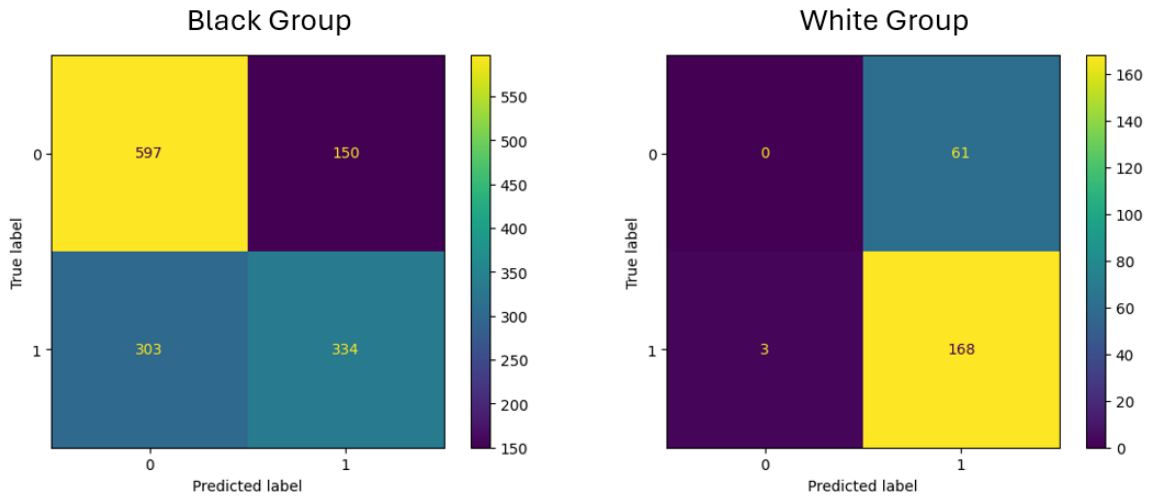

Figure 7: Confusion Matrix of the Race Groups

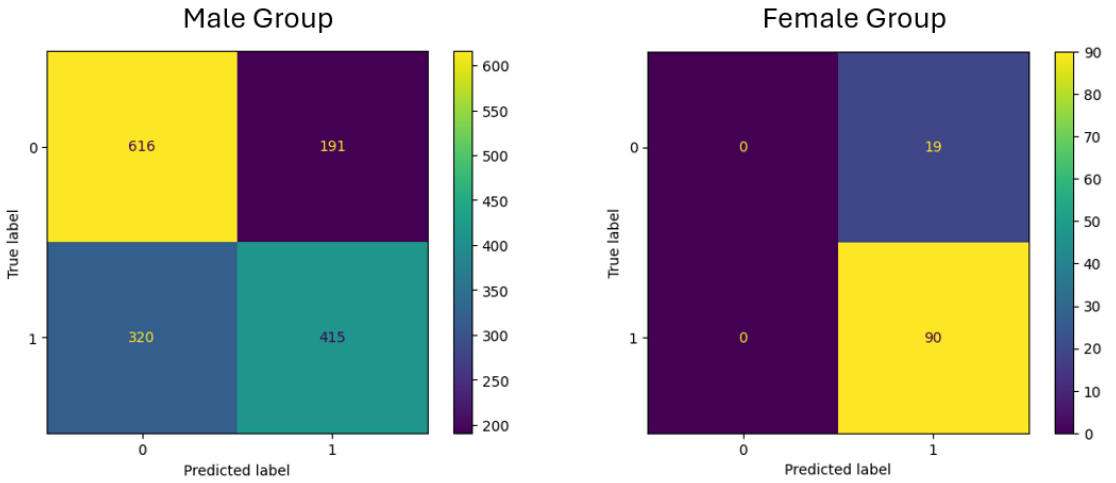

Figure 8: Confusion Matrix of the Sex Groups

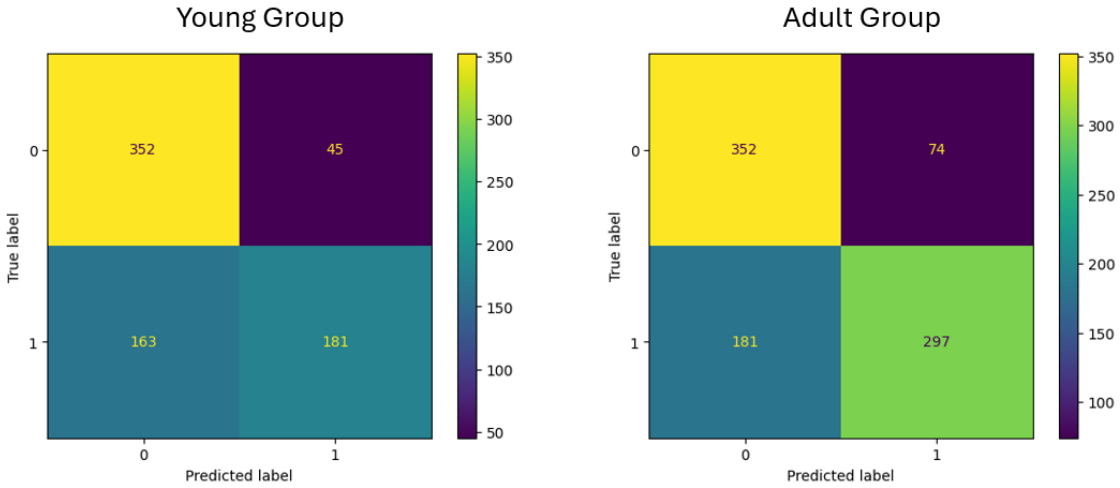

Figure 9: Confusion Matrix of the Age Groups

3.0 Findings and Discussion

The study revealed bias in the model especially within race and sex as shown in the confusion matrix in figures 7, 8, and 9. While total accuracies for the 3 protected characteristics showed no bias, group fairness and equality of opportunity expressed a very high bias within the race and sex groups. The positive rate for Blacks and Males was almost thrice as less as for the Whites and Females in the study while the actual predicted positives proportion when compared is almost twice as less which means Blacks and Males are more exposed to the risk of been murder and could heighten if the victim has both characteristics (a Black male). The study can conclude that there is no bias within the age group or negligible if it exists based on the results in group fairness.

Each fairness criteria have its own limitation. Accuracy can be misleading as it does not consider the trade of in error, that is, no information about False Positive and Negatives which can be unequal among groups and impact results (Rodolfa, Lamba, and Ghani, 2021). Group fairness improves on this but can mask unfair treatment among individuals. Efforts to equalize outcomes across groups can sometimes lead to a reduction in overall model performance, a phenomenon known as "leveling down" (Jui and Rivas 2024). Finally, while equality opportunity is a great criterion as it considers individuality, it does not address disparities in false positive rates. In contexts such as crime and criminal justice, higher false positive rates for certain groups can lead to unjust outcomes which suggests that equality of opportunity may not fully capture the fairness concerns in scenarios where false positives carry significant consequences.

Reference

Dressel, J. and Farid, H., 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1), p.eaao5580.

Hanson, R.F., Sawyer, G.K., Begle, A.M. and Hubel, G.S., 2010. The impact of crime victimization on quality of life. *Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies*, *23*(2), pp.189-197.

https://hmicfrs.justiceinspectorates.gov.uk/publication-html/meeting-the-needs-of-victims-in-the-criminal-justice-system

Jui, T.D. and Rivas, P., 2024. Fairness issues, current approaches, and challenges in machine learning models. *International Journal of Machine Learning and Cybernetics*, *15*(8), pp.3095-3125.

Khademi, A. and Honavar, V., Algorithmic Bias in Recidivism Prediction: A Causal Perspective. arXiv 2019. *arXiv preprint arXiv:1911.10640*.

Mehrotra, D. and Cameron, D., 2023. The maker of ShotSpotter is buying the world's most infamous predictive policing tech. *Wired magazine*.

Montana, E., Nagin, D.S., Neil, R. and Sampson, R.J., 2023. Cohort bias in predictive risk assessments of future criminal justice system involvement. *Proceedings of the National Academy of Sciences*, *120*(23), p.e2301990120.

Rodolfa, K.T., Lamba, H. and Ghani, R., 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, *3*(10), pp.896-904.