2022 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT) | 978-1-6654-6090-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/BDCAT56447.2022.00047

Accuracy-Fairness Tradeoff in Parole Decision Predictions: A Preliminary Analysis

John W. Gardner, Furkan Gursoy, Ioannis A. Kakadiaris
Computational Biomedicine Lab
Dept. of Computer Science
University of Houston
Houston, TX, USA
gardnejw@rose-hulman.edu, {fgursoy, ioannisk}@uh.edu

***Abstract*—Algorithms play an essential and expanding role in public policy decisions, including those in criminal justice. This short paper reports on the first author's summer research project characterizing the tradeoff between accuracy and fairness in parole decision predictions. The dataset employed in this study contains over 30,000 parole decisions made by the New York State Division of Criminal Justice Services. Each decision contains information on the subject, such as sex, race/ethnicity, and parole decision, as well as predictive features describing the crime committed by the subject and the parole interview held. Logistic regression, decision tree, support vector machine, and random forest models are trained and utilized to analyze parole decision predictions based on the available features. Most models fail to pass standard fairness tests for most fairness metrics. Moreover, while there may be an overall tradeoff between fairness and accuracy, the obtained differences in accuracy are too small to make a well-supported claim. Future research may enhance the preliminary work introduced in this paper by using multiple real-world datasets to investigate the tradeoff between accuracy and fairness.**

***Index Terms*—Machine Learning, Parole Decisions, Accuracy, Fairness, Tradeoff**

I. Introduction

Artificial intelligence and machine learning (AI/ML) play a crucial role in decision-making for public policy [1]. They play an increasingly essential role in many branches of government in the United States and worldwide. Meanwhile, standards for accountability and societal concerns have fallen behind the use and influence of AI/ML.

Two concerns relating to critical decision-aiding systems, such as those in the criminal justice system, are accuracy and fairness. Accuracy concerns whether an AI/ML model has acceptable levels of predictive accuracy so that the errors made by the model and, consequently, the potentially erroneous decisions are minimized. Fairness, in general, is concerned with whether the decisions from an AI/ML are fair across different groups or individuals, usually based on sensitive group membership such as race or sex. However, tradeoffs usually exist between fairness and accuracy [2].

Using a real-world parole decision dataset from the state of New York, this work aims to identify and characterize any tradeoff between accuracy and fairness for the machine learning-based parole decision prediction models.

II. Methodology

A. Dataset and Preprocessing

The Parole Hearing dataset released by the New York State Parole Board is utilized to analyze the potential tradeoff between fairness and accuracy. The dataset was initially scraped from the Parole Hearing Data Project repository as described on the project's GitHub page [3]. As the original dataset was taken off the internet, the dataset used in this paper has instead been pulled from a copy available on Kaggle [4]. The dataset contains parole hearings from 2012 to 2016. In addition to parole interview decision, the available features are:

- The housing or interview facility: where the subject in question is housed or interviewed for parole (one of 69 possible locations)
- The crime of conviction: the type of crime the subject committed to sentence the subjects time in prison (one of more than 200 types of crime such as animal fighting, bail jumping, bribery, grand larceny, identity theft, manslaughter, robbery, stalking, welfare fraud)
- The class of crime: the class of the crime the subject committed (A, B, C, D, or E based on the maximum term of imprisonment for the offense)
- The parole interview type: the type of interview the subject is given for parole (11 different types such as initial, merit time, reappear, and medical)
- Race/ethnicity (American Indian/Alaska Native, Asian/Pacific, Black, Hispanic, Other, Unknown, and White)
- Sex (Female, Male)
- Crime count: the number of crimes the subject has been charged for

The target variable, parole interview decision, may have one of several decisions as its value. The decisions are encoded as binary by assigning decisions such as "granted" and "paroled" to the positive class and decisions such as "denied" and "not granted" to the negative class. Instances with missing data or unclear decisions such as "or earlier/postponement" and "rescind original release date/new date" are removed.

To allow different machine learning models to train and test on the dataset, categorical variable encoding methods were utilized for categorical variables. Target encoding replaces a

978-1-6654-6090-3/22/$31.00 ©2022 IEEE
DOI 10.1109/BDCAT56447.2022.00047

categorical value with the average target value of all instances in the same category [5]. Target encoding was used for the housing or interview facility, the crime of conviction, the parole interview type, and the class of crime. Another encoding method, *k-1* encoding, which represents each categorical value with a new variable except for the reference value, was also utilized. Specifically, *k-1* encoding was utilized for the race/ethnicity feature using white as the reference group. The sex feature was also processed to indicate female with 0 and male with 1.

The distributions of the target variable and the sensitive group memberships are as follows (with corresponding frequencies provided inside parentheses) Parole Decision: No Parole (25,082), Parole (7,849); Ethnicity/Race: American Indian/Alaskan (362), Asian/Pacific (162), Black (14,269), Hispanic (6,213), Other (390), Unknown (433), White (11,102); Sex: Male (30,878), Female (2,053).

B. Machine Learning Models

Several machine learning models were implemented to analyze the New York State Parole dataset. Specifically, logistic regression [6], decision tree [7], random forest [8], and support vector machine [9] were used.

As no training/test data split was initially provided, the dataset was randomly split into 80/20 training/test sets. For all models, the same learning procedure was followed. First, the best hyperparameter configuration was found via a grid search strategy employing a 5-fold cross-validation on the training set. The models were then trained on the whole training set with their best hyperparameter configurations. Finally, the trained models were analyzed for their accuracy and fairness performance on the test set.

C. Evaluation

Accuracy. Table I describes the confusion matrix for binary classification tasks, as is the task in this paper. The four cells denote true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) based on whether predictions are true for each class of actual values.

TABLE I
CONFUSION MATRIX FOR BINARY CLASSIFICATION

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Based on the confusion matrix, the following metrics can be computed to assess predictive accuracy.

- Accuracy— (TP + TN) / (TP + FN + TN + FP)
- Precision— TP / (TP + FP)
- Recall— TP / (TP + FN)
- F1-score— (2 * Precision * Recall) / (Precision + Recall)

All scores range from 0 to 1, with 1 indicating the perfect score. A perfect accuracy indicates that all predictions were correct. A perfect precision indicates that all positive predictions were positive. A perfect recall indicates that all actual positive classes were correctly predicted as positives. F1-score is the harmonic mean of precision and recall. While accuracy provides a general view of predictive accuracy, other metrics offer more insights into the types of errors made.

Fairness. Fairness in machine learning concerns the unjust treatment of groups or individuals, usually based on sensitive characteristics. This paper focuses on group fairness based on race and sex. Aequitas [10], a fairness audit toolkit for classification tasks, is employed to analyze fairness. The available disparity tests are for the following metrics: predicted positive rate (PPR), predicted positive group rate (PPGR), false positive rate (FPR), false discovery rate (FDR), false negative rate (FNR), and false omission rate (FOR). For each metric, the scores are desired to be on par across different sensitive groups. The test results are with respect to a reference group. For instance, a value of 1.3 for a specific group and a metric indicates that the said group has a 30% larger value for the said metric relative to the reference group. Detailed descriptions can be found in the tool's original documentation [10].

III. RESULTS

A. Accuracy

Table II presents the predictive accuracy results for each model. Metrics are computed by treating parole and no parole decisions separately as the positive class. The accuracy metric does not depend on the positive class and thus has the same value in both cases. Overall, all models have similar predictive performance. An observation is that the algorithms achieve a better performance for no parole cases in comparison to parole cases. This may be due to the fact that the data set includes larger number of no parole decisions than parole decisions. In general, the random forest model performs slightly better than the other models, followed by the logistic regression, decision tree, and support vector machine models. Such slight differences imply that the available predictor features may not be informative enough. Alternatively, the dataset, as is, may not have complicated relationships between predictor variables and the target variable. Hence the problem does not require more complex models such as random forests to learn better.

B. Fairness

The results from the Aequitas bias and fairness audit toolkit are reported in Tables III, IV, V, and VI for the logistic regression, decision tree, random forest, and support vector machine models, respectively. For race and ethnicity, the reference baseline group is set as white. For sex, the reference group is set as female. The values lower than 0.80 or greater than 1.25 are generally considered problematic [11], whereas the value of 1 corresponds to the perfect parity with the reference group, i.e., the group has the same value as the white group for the corresponding metric.

The fairness results in Tables III, IV, V, and VI rarely meet the standard criteria for fairness tests, i.e., the figures are not usually within the 0.80 to 1.25 range. For instance, in Table III, the black group has a PPR parity score of approximately half the same as the white group. Similarly, the black group's

TABLE II
ACCURACY RESULTS

Model	Positive Class	Precision	Recall	F1-score	Accuracy
Logistic Regression	No parole	0.78	0.97	0.86	0.7692
	Parole	0.59	0.14	0.22	
Decision Tree	No parole	0.78	0.96	0.86	0.7686
	Parole	0.57	0.15	0.24	
Random Forest	No parole	0.78	0.97	0.87	0.7712
	Parole	0.61	0.13	0.22	
Support Vector Machine	No parole	0.78	0.96	0.86	0.7660
	Parole	0.57	0.15	0.24	

PPGR, FPR, and FOR parity scores are also low. On the other hand, the parity between the black and white groups is satisfied in terms of FDR and FNR metrics, as scores for those metrics fall within the acceptable range.

There are also some outlier values, such as zero values for the American Indian/Alaska Native group in Table V for FDR and FPR metrics or very large values for the Asian/Pacific group in Table VI for the FPR metric. Such large deviations are caused by the lower number of instances in those groups since the smaller sample size increases the variance of statistics.

To compare the overall fairness performances of the models, average statistics are computed for each model across all fairness metrics. Considering the outlier cases in certain groups due to the small available sample sizes, a weighted average computation strategy is utilized. First, the absolute distance between the metric value and the reference group value (a reference group's value is always 1 by definition) is computed for each metric. Then, a weighted average is computed over sensitive groups using corresponding group sample sizes as the weights. The resulting values represent the average distances from the perfect parity with respect to the corresponding models and fairness metrics. The results are provided in Table VII for race and in Table VIII for sex. The best values (i.e., minimum distances from the perfect parity) are shown in bold.

When models are compared for their fairness performance based on the weighted average distance metric, no model is consistently the best for all metrics or all types of sensitive categories (i.e., both based on race or sex). Simpler models such as decision tree and logistic regression rank better for fairness. The former obtains better fairness scores for race-based groups, while the latter obtains better fairness scores for sex-based groups.

TABLE III
LOGISTIC REGRESSION FAIRNESS AUDIT RESULTS (PARITY SCORES)

	Group	PPR	PPGR	FDR	FPR	FOR	FNR
Race/ Ethnicity	Amer Ind/Alsk	0.01	0.17	0.00	0.00	0.86	1.12
	Asian/Pac	0.13	8.43	0.84	14.98	2.77	0.46
	Black	0.53	0.42	0.97	0.37	0.77	1.08
	Hispanic	0.49	0.88	0.93	0.79	0.90	1.00
	Other	0.06	1.62	0.69	1.11	0.82	0.81
	Unknown	0.01	0.14	0.00	0.00	1.04	1.14
	White	1.00	1.00	1.00	1.00	1.00	1.00
Sex	Female	1.00	1.00	1.00	1.00	1.00	1.00
	Male	1.91	0.13	0.99	0.10	0.71	1.66

TABLE IV
DECISION TREE FAIRNESS AUDIT RESULTS (PARITY SCORES)

	Group	PPR	PPGR	FDR	FPR	FOR	FNR
Race/ Ethnicity	Amer Ind/Alsk	0.02	0.47	0.77	0.34	0.84	1.07
	Asian/Pac	0.05	3.20	0.77	5.25	2.64	0.89
	Black	0.76	0.60	1.09	0.60	0.76	1.05
	Hispanic	0.51	0.91	0.87	0.76	0.89	0.98
	Other	0.03	0.75	0.93	0.69	0.98	1.04
	Unknown	0.02	0.40	1.55	0.62	1.08	1.15
	White	1.00	1.00	1.00	1.00	1.00	1.00
Sex	Female	1.00	1.00	1.00	1.00	1.00	1.00
	Male	1.73	0.12	0.87	0.08	0.70	1.82

TABLE V
RANDOM FOREST FAIRNESS AUDIT RESULTS (PARITY SCORES)

	Group	PPR	PPGR	FDR	FPR	FOR	FNR
Race/ Ethnicity	Amer Ind/Alsk	0.01	0.19	0.00	0.00	0.86	1.10
	Asian/Pac	0.11	7.12	0.80	12.04	2.52	0.61
	Black	0.71	0.56	1.11	0.57	0.76	1.05
	Hispanic	0.53	0.95	0.84	0.77	0.89	0.98
	Other	0.04	1.13	1.28	1.42	0.99	1.01
	Unknown	0.01	0.33	1.28	0.42	1.05	1.12
	White	1.00	1.00	1.00	1.00	1.00	1.00
Sex	Female	1.00	1.00	1.00	1.00	1.00	1.00
	Male	1.43	0.10	0.92	0.07	0.76	1.89

TABLE VI
SUPPORT VECTOR MACHINE FAIRNESS AUDIT RESULTS (PARITY SCORES)

	Group	PPR	PPGR	FDR	FPR	FOR	FNR
Race/ Ethnicity	Amer Ind/Alsk	0.04	1.06	1.19	1.19	0.82	0.99
	Asian/Pac	0.16	10.30	0.84	18.39	2.58	0.58
	Black	0.47	0.37	1.02	0.34	0.76	1.06
	Hispanic	0.49	0.87	0.94	0.78	0.91	1.00
	Other	0.10	2.56	1.19	3.01	0.90	0.86
	Unknown	0.08	1.82	1.19	2.16	0.96	0.93
	White	1.00	1.00	1.00	1.00	1.00	1.00
Sex	Female	1.00	1.00	1.00	1.00	1.00	1.00
	Male	0.94	0.06	0.72	0.04	0.66	1.58

TABLE VII
AVERAGE MODEL FAIRNESS FOR RACE AND ETHNICITY
(WEIGHTED AVERAGE STATISTIC)

Model	PPR	PPGR	FDR	FPR	FOR	FNR
Logistic Regression	0.51	0.52	0.09	0.62	**0.20**	0.07
Decision Tree	**0.36**	**0.33**	0.12	**0.39**	**0.20**	**0.04**
Random Forest	0.39	0.38	0.15	0.47	**0.20**	0.05
Support Vector Machine	0.55	0.57	**0.04**	0.70	**0.20**	0.05

TABLE VIII
AVERAGE MODEL FAIRNESS FOR SEX
(WEIGHTED AVERAGE STATISTIC)

Model	PPR	PPGR	FDR	FPR	FOR	FNR
Logistic Regression	0.91	**0.87**	**0.01**	**0.90**	0.29	0.66
Decision Tree	0.73	0.88	0.13	0.92	0.30	0.82
Random Forest	0.43	0.90	0.08	0.93	**0.24**	0.89
Support Vector Machine	**0.06**	0.94	0.28	0.96	0.34	**0.58**

C. Accuracy Fairness Tradeoff

While a tradeoff between accuracy and fairness is generally existent, it is not fully obvious due to the small differences in observed accuracy values. For the race and ethnicity grouping, predictive accuracy (see Table II) and fairness performance (based on being a top model as shown in bold in Table VII) are compared as follows.

- The random forest model has an accuracy of 77.12% and is a top choice in only one of the parity tests.
- The logistic regression model has an accuracy of 76.92% and is a top choice in only one of the parity tests.
- The decision tree model has an accuracy of 76.86% and is a top choice in five parity tests.
- The support vector machine model has an accuracy of 76.60% and is a top choice in two parity tests.

The findings are similar for the sex-based groups, as can be observed in Table VIII.

- The random forest model is a top choice in only one of the parity tests.
- The logistic regression model is a top choice in three of the parity tests.
- The decision tree model is never a top choice in any of the parity tests.
- The support vector machine model is a top choice in two of the parity tests.

While not perfectly consistent, models with better predictive accuracy (e.g., random forest) have relatively worse outcomes in the fairness audit.

IV. CONCLUSION

This short paper aims to characterize the tradeoff between accuracy and fairness in parole decision predictions. All machine learning-based models employed in this study obtained similar predictive performances but varied in their performance in different fairness metrics. The tradeoff between accuracy and fairness is not very obvious. However, the results hint at a tradeoff where improved accuracy may come at the cost of reduced fairness. When predictive performance is comparable across different models, as with the models in this work, fairer models can be preferred over other slightly more accurate models.

This work serves as a preliminary analysis and has many limitations. Specifically, the limitations of this study include the following. First, only a single dataset from a specific region is utilized. Second, judging by the obtained predictive accuracy, the dataset does not appear to contain all relevant variables for a parole decision. This limitation also hindered obtaining any meaningful accuracy differences between different models. Third, the binarization of the target variable may be too simplistic. Therefore, rather than presenting conclusive evidence, this paper motivates future work to further investigate the research question tackled in this work.

ACKNOWLEDGMENT

The first author's work was supported by the University of Houston's Computer Science REU program that is primarily sponsored by the National Science Foundation under Award CCF-195029 and the University of Houston's College of Natural Sciences and Mathematics. Drs. Gursoy and Kakadiaris' work was supported by the National Science Foundation under Award CCF-2131504. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] J. Monahan and J. L. Skeem, "Risk redux: The resurgence of risk assessment in criminal sanctioning," *Federal Sentencing Reporter*, vol. 26, no. 3, pp. 158–166, 2014.
[2] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3-44, 2021.
[3] N. Zeichner, R. Ackerman, J. Krauss, J. Adams "The Parole Hearing Data Project," 2019, https://github.com/rcackerman/parole-hearing-data (retrieved Aug. 25, 2022).
[4] Parole Hearing Data Project, "Parole hearings in New York State," 2016, https://www.kaggle.com/datasets/parole-hearing-data/parole-hearings-in-new-york-state (retrieved Aug. 25, 2022).
[5] M. Halford, "Target encoding done the right way," 2018, https://maxhalford.github.io/blog/target-encoding/ (retrieved Aug. 25, 2022).
[6] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B*, vol. 20, no. 2, pp. 215-232, 1958.
[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, New York, NY, USA: Routledge, 1983.
[8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2004.
[9] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large-Margin Classifiers*, vol. 10, no. 3, pp. 61-74, 1999.
[10] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R.Ghani, "Aequitas: A bias and fairness audit toolkit," 2018, https://arxiv.org/abs/1811.05577 (retrieved Aug. 25, 2022).
[11] T. A. Stetz, "The 4/5ths rule," in T*est Bias in Employment Selection Testing*, Cham, Switzerland: Springer, 2022, pp. 67–74.