Document Title:	NIJ Recidivism Forecasting Challenge Report for Team PASDA
Author(s):	Michael Porter, George Mohler
Document Number:	305042
Date Received:	July 2022
Award Number:	NIJ Recidivism Forecasting Challenge Winning Paper

NIJ Recidivism Forecasting Challenge Report for Team PASDA

Michael Porter and George Mohler

1 Introduction

The 2021 NIJ recidivism forecasting challenge is a competition hosted by the National Institute of Justice with the aim to "increase public safety and improve the fair administration of justice across the United States"[1]. The challenge focuses on data from the State of Georgia on individuals released from prison to parole supervision for the period January 1, 2013, through December 31, 2015. Challenge participants are tasked with constructing a predictive model of 1, 2, and 3 year recidivism upon release from prison based on variables such as age, gender, race, education, prior arrests and convictions, and other covariates.

The scoring metric used in a majority of categories of the competition is the mean square error (Brier score):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2, \tag{1}$$

where y_i is the binary recidivism outcome for individual i indicating recidivism $(y_i = 1)$ or no recidivism $(y_i = 0)$, p_i is the forecasted probability of recidivism, and N is the number of individuals in the dataset. Given recent concerns of bias of predictive models of recidivism, such as disparate false positive and negative rates across different racial/ethnic groups [3, 5], the NIJ challenge includes a second set of categories aimed at balancing low MSE while reducing the difference of false positive rates (denoted FP below) between groups of Black and white individuals in the data. In particular, contestants' models are scored according to the metric:

$$(1 - MSE)(1 - |FP_{Black} - FP_{white}|). \tag{2}$$

We refer to this metric as the "NIJFM" (NIJ Fairness Metric). False positive rates require a binary prediction defined by a cutoff, which in the NIJ competition is defined to be $p_i \geq 0.5$. Whereas the goal in the first set of categories is to minimize MSE, the goal in the second set of categories is to maximize the NIJFM (which occurs when the MSE and the difference of false positive rates are close to zero). Unlike a number of loss functions in the fairness-aware machine learning literature that have additive penalties to encourage some

[1] https://nij.ojp.gov/funding/recidivism-forecasting-challenge

This resource was prepared by the author(s) using Federal funds provided by the U.S.
Department of Justice. Opinions or points of view expressed are those of the author(s) and do not
necessarily reflect the official position or policies of the U.S. Department of Justice.

type of fairness [1, 2], the NIJFM in Equation 2 is a multiplicative loss function defined by the product of the target loss (MSE) and the fairness penalty (difference in false positive rates). In this short note we explore several regression based methods for optimizing the NIJFM using data from the NIJ competition.

2 Relevant Literature

There are several fairness-aware methods that have been introduced in the literature for forecasting recidivism. In [1, 2], the authors consider a convex surrogate loss where the step function representing the decision at the cutoff is replaced by a linear approximation (simply the score itself):

$$MSE + \lambda \Big(\sum_{X_i \in S_{00}} \frac{X_i^t \theta}{|S_{00}|} - \sum_{X_i \in S_{10}} \frac{X_i^t \theta}{|S_{10}|} \Big)^2 . \tag{3}$$

Here S_{00} is the set of individuals of race 0 that did not recidivate $(y_i = 0)$ and S_{10} is the set of individuals of race 1 that did not recidivate. The penalty term encourages the average scores over the negative class $(y_i = 0)$ to be matched across race (as λ increases). This is a form of group fairness where we wish false positive rates to match across groups (alternatively individual fairness can be defined by bringing the summation outside of the squared term [2]). Because the loss function in Equation 3 is quadratic, there is an analytical solution determined by the linear system:

$$\Big[\frac{2}{N} X^t X + 2\lambda (V_0^t V_0 - V_0^t V_1 - V_1^t V_0 + V_1^t V_1) \Big] \theta - \frac{2}{N} X^t y = 0, \tag{4}$$

where $V_j = \sum_{X_i \in S_{j0}} \frac{X_i^t}{|S_{j0}|}$.

Fairness can also be encouraged by post-processing the scores [12]. In the competition we used a post-processing technique for the female category that forces the false positive rates to zero by truncating all scores to the cutoff value (minus 0.0001) if they are above the cutoff. We refer to this method as linear regression with truncation.

3 Variables

3.1 Were variables added to the data set? If so, detail the variables.

We only utilized variables provided by NIJ for the competition.

3.2 What variables were constructed? How were the variables constructed?

We converted logical features into binary, encoded all ordinal features as integers (starting at 0), and left nominal features as is to allow each model to use separately (e.g., dummy coding for linear models, one-hot for xgboost, target-based for catboost). We created several new features. Total prior arrests, total prior convictions, total violations, and total percent of positive drug tests were created by adding the integer encoded ordinal or logical values from the corresponding raw features. Due to the censoring in the ordinal features (e.g., prior felonies is capped at "10 or more" and prior gun charges are reported as either 0 or "more than 1"), these are not expected to be true totals, but rather a simple way to combine information from potentially similar features. We created additional features by standardizing these totals by the (integer encoded) individual's age at release (plus 1) to give a simple estimate of the average events per year. Other created features include: the proportion of programs attended, the average number of drug tests per year (365/average days between drug tests), and the difference between the percentage of days employed and jobs per year. We discovered one perfect predictor possibly due to leakage: a missing value for percent days employed or jobs per year indicated no recidivism. It didn't occur very often in the data (534 times in training and 274 times in the evaluation data), so we encoded it as a binary feature and in a post-processing step converted all final scores with this feature to 0. We also utilized ID as a feature by treating it as time and smoothing the target recidivism variable with respect to ID.

Eleven features contained missing values. For the nominal features, we treated missing values as another level. We overlooked that there were missing values in the supervision risk scores; these received an integer value of 10. Mean imputation was used for missing percent of positive drug tests and median imputation was used for missing average days per drug test. The training and evaluation data were imputed separately.

3.3 Which variables were statistically significant?

In Table 1 we display feature importances from the top performing Catboost model. Several of the hand-crafted features were top predictors including total arrests normalized by release age and the difference between the percentage of days employed and jobs per year. We also observe that the label smoothed over ID (Yavgid) is a top performing predictor variable in the Catboost model.

3.4 What variables were not statistically significant? How was this handled? For example, were they dropped from the overall model?

Catboost does not yield statistical significance estimates for features, however in Table 2 we display p-values of coefficients in a linear regression of recidivism. P-values less than 0.05 are highlighted in bold. One

Feature	Catboost Importance
Arrests_Age	1.23E+01
day_job_diff	7.36E+00
Percent_Days_Employed	5.67E+00
Jobs_Per_Year	5.20E+00
Convictions_Age	4.87E+00
Yavgid	4.67E+00
Gang_Affiliated	4.28E+00
num_DrugTest	3.49E+00
Total_Arrests	3.27E+00
Violations_Age	3.14E+00
Avg_Days_per_DrugTest	2.86E+00
Supervision_Risk_Score_First	2.77E+00
Arrests_Drug	2.50E+00
Dependents	2.31E+00
DrugTests_Meth_Positive	2.18E+00
DrugTests_THC_Positive	2.08E+00
Total_Convictions	2.05E+00
Gender	2.01E+00
Arrests_PPViolationCharges	1.80E+00
Arrests_Property	1.79E+00
Residence_Changes	1.68E+00
DrugTests_Age	1.48E+00
Program_Attendances	1.34E+00
DrugTests_Cocaine_Positive	1.33E+00
Supervision_Level_First	1.32E+00
Total_DrugTests_Positive	1.29E+00
Convictions_Misd	1.05E+00
Revos_Parole	1.02E+00
Convictions_Drug	1.02E+00
Prison_Years	9.97E-01
Delinquency_Reports	9.14E-01
prop_Program_Attendance	9.14E-01
NA Indicator Days Employed/Jobs Per Year	9.02E-01
Prison_Offense	8.18E-01
Arrests_Felony	7.90E-01
Age_at_Release	7.69E-01
Education_Level	7.41E-01
Convictions_Felony	6.53E-01
Arrests_Misd	5.11E-01
Violations_Instruction	4.92E-01
Total_Violations	4.42E-01
DrugTests_Other_Positive	3.93E-01
Condition_Cog_Ed	3.61E-01
Convictions_PPViolationCharges	3.54E-01
Arrests_Violent	2.49E-01
Employment_Exempt	2.45E-01

Table 1: Feature importance from Catboost model (importance greater than .02).

Variable	p-val	Variable	p-val	Variable	p-val	Variable	p-val
(Intercept)	0.63015	Arrests_Felony	0.71025	PUMA17	0.14044	Condition_Other	0.10921
Gender	$< \mathbf{10^{-10}}$	Arrests_Misd	0.53004	PUMA18	0.39863	Viol_ElectMonit	0.11626
Race	0.722	Arrests_Violent	0.9729	PUMA19	0.31207	Viol_Instr	**0.01255**
Age_at_Rel	**0.0023**	Arrests_Property	0.05677	PUMA20	0.09028	Viol_FailToReport	0.71921
PUMA1	0.91754	Arrests_Drug	0.55387	PUMA21	0.59454	Viol_MoveWoutPerm	0.1901
PUMA2	0.98694	Arrest_PPViolChg	0.1953	PUMA22	0.21108	Delinquency_Reports	0.2891
PUMA3	0.68281	Arrests_DVCharges	0.29658	PUMA23	0.79775	Program_Attendances	0.18822
PUMA4	0.31634	Arrests_GunCharges	0.27443	PUMA24	0.38711	Prog_UnexAbs	0.2234
PUMA5	0.3171	Conv_Felony	0.96811	Gang_Affilfalse	$< \mathbf{10^{-8}}$	Residence_Changes	0.15151
PUMA6	0.70927	Conv_Misd	**0.02789**	Super_Risk_First	0.07341	Avg_Days_per_DT	0.05445
PUMA7	0.81599	Conv_Viol	0.81435	Super_Level_First	0.98807	DT_THC_Positive	0.88352
PUMA8	0.26577	Conv_Prop	0.8603	Educ_Level	0.39698	DT_Cocaine_Positive	0.40606
PUMA9	0.27584	Conv_Drug	0.14056	Dependents	**0.00276**	DT_Meth_Positive	**0.0048**
PUMA10	0.10515	Conv_PPViolChg	0.48432	Pris_OffDrug	0.38997	DT_Other_Pos	0.60194
PUMA11	0.19515	Conv_DomViolChg	0.6987	PrisViol.Non.Sex	0.46816	Perc_Days_Employed	$< \mathbf{1-^{-6}}$
PUMA12	0.30476	Conv_GunCharges	0.92369	Prison_OffProp	0.68044	Jobs_Per_Year	$< \mathbf{10^{-5}}$
PUMA13	0.39027	Revos_Parole	**0.03281**	Prison_OffenseNA	0.53394	Employment_Exempt	0.74186
PUMA14	0.98969	Revos_Prob	0.33358	Pris_OffOther	0.44563	prop_Prog_Att	0.117
PUMA15	0.61716	Condition_MH_SA	**0.01298**	Pris_Yr	0.21824	num_DrugTest	0.50401
PUMA16	0.62571	Condition_Cog_Ed	0.06247	Arrests_Age	0.06886	Yavgid	0.22772

Table 2: P-values of variables in linear regression of recidivism.

advantage of boosting is that it provides some natural feature selection, so that we did not have to explicitly drop variables from the model (although doing so could have improved the performance, e.g. some form of backward or forward selection).

4 Models

4.1 What type of model was used?

We considered several different model families throughout the duration of the contest. These included: unpenalized linear models, penalized linear models (i.e., lasso, ridge, elasticnet, and relaxed lasso), generalized additive models (GAM), boosted trees (GBM, xgboost, catboost), and bagged trees (random forest). Select interaction effects were considered in the linear models. All model parameters were fit using squared error loss and predictions were truncated to be between 0 and 1. For the penalized linear models, ten-fold cross-validation was used to compute the penalty parameter(s). GAM models use AIC to set the smoothing levels. For the tree-based models, very limited model tuning was performed. At the start of each round, the tuning parameters were set by using a small subset of the training data to perform a grid search over a coarse grid of tuning parameters.

A form of stacking was used to combine predictions from the individual models. Specifically, we first made out-of-sample predictions from all individual models. These predictions are treated as additional features and used to fit an ensemble model on the out-of-sample data. We considered several stacking models including: unpenalized linear models, penalized linear models (i.e., lasso, ridge, elasticnet, and relaxed lasso), generalized additive models (GAM), and best subsets. The performance on the stacking ensemble models were evaluated on another out-of-sample data set. Interactions with gender, race, and age at released were considered. Squared error loss was used to estimate the stacking weights.

A summary of our method's performance on the evaluation data is given in Table 3. As the base recidivism rate deceases each round, we see our performance correspondingly increasing. While the truncation to the decision threshold produces the desired ΔFPR = 0, the unadjusted forecasts (Men in rounds 1 and 2) had a small ΔFPR indicating that our models didn't produce a large racial difference in false positive rates. More detailed results are provided in the Appendix.

Table 3: Performance of our forecasts on the evaluation data. Recidivism is the recidivism rate for the round.

Round	Gender	n	Recidivism	MSE	ΔFPR	AF
1	F	950	21.9%	0.1554	0.0000	0.8446
	M	6857	31.2%	0.1915	0.0067	0.8032
2	F	742	17.7%	0.1245	0.0000	0.8755
	M	4718	25.1%	0.1638	0.0047	0.8323
3	F	611	15.1%	0.1165	0.0000	0.8835
	M	3535	20.7%	0.1522	0.0000	0.8478

A description of the base models used for stacking in each round is given in Table 4. Each base model was fit a number of times using a different random seed to get a diversity of estimates. For each seed, an ensemble was created by averaging the predictions from all models. The final forecast was obtained by averaging the ensemble estimates over all seeds and truncating to $\hat{p}_i \in [0, .50)$ where specified. For all rounds we found that boosted trees were favored. The model CatBoost Ensemble is an ensemble of four slightly over-fit CatBoost models using different tree depths.

We tried several different ways to estimate the stacking weights, but found that an equally weighted average of a small number of base models performed best using our cross-validation scheme. All possible combinations of (up to 15) base models were considered for the ensembles in each round. The small number of base models selected follows the conclusions of [4]. While we ended up using different ensembles for males and females (with the exception of round 1), the performance estimates were similar for both groups using either model.

Most base models we considered are stochastic and sensitive to the random seed that controls the internal resampling and other aspects; e.g., boosted and bagged tree models use bootstrap or sub-sampled data to create each tree in the ensemble as well as the random set of features that are considered at each split and

Table 4: Description of base models used for stacking. Each base model was fit # seeds times using a different random seed. Trunc indicates if the ensemble estimates were truncated to $[0, 0.50)$.

Round	# seeds	Gender	Trunc	Models
1	100	F	Yes	1. CatBoost (depth = 4) 2. CatBoost (depth = 6) 3. Linear Regression (hand-selected features) 4. Ridge Regression
		M	No	1. CatBoost (depth = 4) 2. CatBoost (depth = 6) 3. Linear Regression (hand-selected features) 4. Ridge Regression
2	200	F	Yes	1. CatBoost Ensemble (depths = 5,6,7,8) 2. XGBoost (max.depth = 5)
		M	No	1. CatBoost (depth = 8) 2. CatBoost Ensemble (depths = 5,6,7,8) 3. XGBoost (max.depth = 5)
3	200	F	Yes	1. CatBoost Ensemble (depths = 3,4,5,6) 2. Relaxed Lasso ($\gamma = 1/2$)
		M	Yes	1. CatBoost Ensemble (depths = 3,4,5,6) 2. XGBoost (max.depth = 3)

the penalized regression models use cross-validation to select the penalty strength. Figure 1 shows the range (difference between maximum and minimum) of the (pre-truncated) probability estimates for round 2 as a function of the average probability. The values are based on 200 model fits with different random seeds. This indicates that the variability in predicted probabilities increases as a function of the average probability increases and is most pronounced above the 0.50 decision threshold where the average range is in excess of 7% for males and 10% for females. The range for females is larger than males which we speculate is due to a two-component ensemble used for females but three-component ensemble used for males.

Figure 2 shows the MSE distribution over the seeds. The variability of the female MSE values is larger than that of males. If we used the predictions from a single seed our MSE score could differ by more than .0016. The MSE obtained from averaging predictions over seeds is slightly better than the average MSE. We did not try any other methods of combining the predictions over seeds, but its possible that a bumping [11] type approach that uses the model from the single best seed could lead to improved performance. We truncated the female scores in round 2 according to the average probability, but could have considered other aspects like the proportion of estimates that exceeded the threshold to adjust the estimated probabilities to better address fairness criteria.

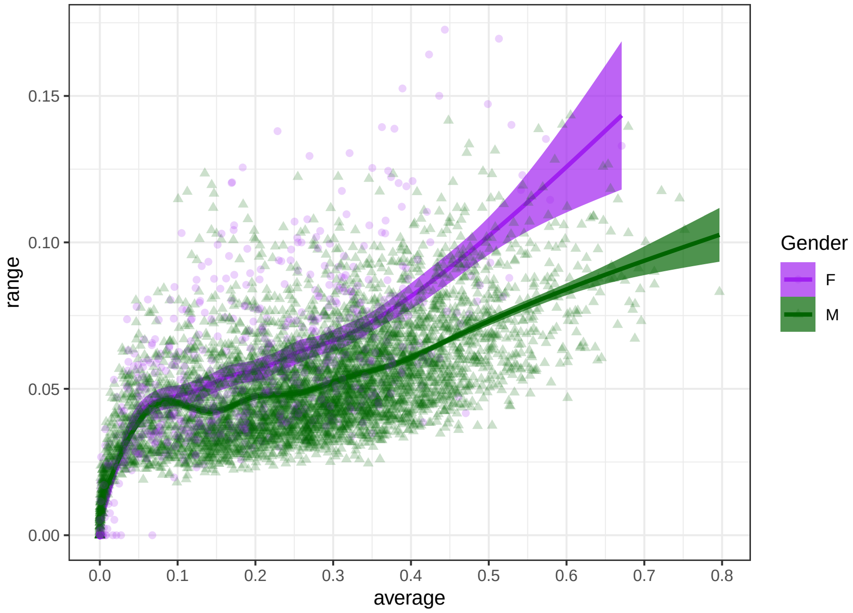

Figure 1: The range of pre-truncated predicted recidivism probabilities as a function of the average probability for round 2. The values are based on 200 model fits with different random seeds.

4.2 Did you try other models? Were they close in performance? Not at all close?

We tried other models such as feedforward neural networks, support vector machines, and random forest. These models yielded reasonably accurate predictions, however they slightly decreased the ensemble model therefore we did not include them in the submission.

5 Future Considerations

5.1 What other evaluation metrics should have been considered/used for this Challenge?

While the MSE can be used to estimate recidivism forecasting models, other metrics may be better suited for model evaluation. In practice, confusion tables that contain accuracy, false positive rates, and false negative rates will better highlight the tradeoffs between different models and cutoff choices. Given that results can change depending on the choice of cutoff, comparing cost, ROC, and precision-recall curves (see Figure 3) may provide more insight than a single metric. Also, the type and severity of the crime committed could be incorporated into a metric, similar to how the gini index is used in evaluating insurance risk models [7]. Finally, we note that there are alternative definitions of fairness (such as individual fairness) that have been

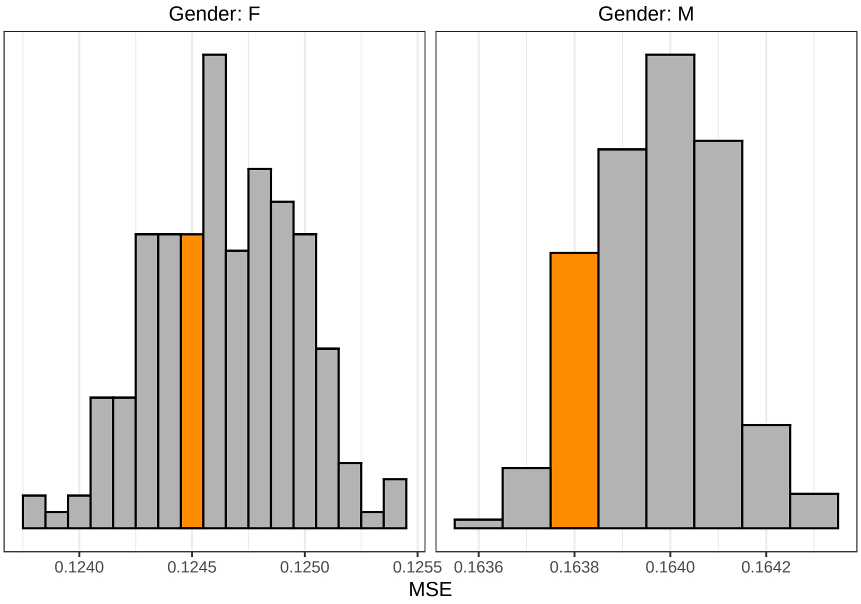

Figure 2: Distribution of MSE scores from different seeds in round 2. The orange bar is the bin containing the MSE we obtained using the average probability.

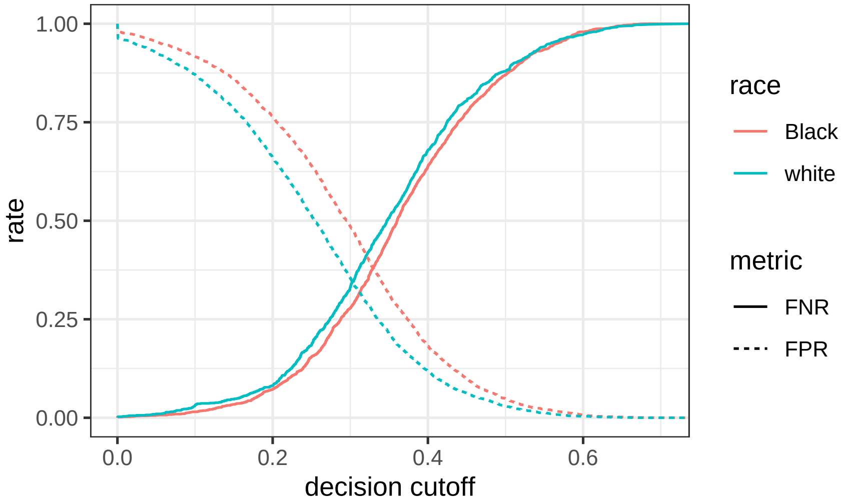

Figure 3: False positive and negative rate vs. decision cutoff by race for linear regression.

discussed in the literature [9, 6], and these may provide a more nuanced assessment of recidivism forecasts than group false positive rates.

Model	MSE	FP (Black)	FP (white)	NIJFM
Linear Reg.	0.192 (0.002)	0.048 (0.004)	0.030 (0.003)	0.793 (0.005)
Logistic Reg.	0.192 (0.002)	0.067 (0.004)	0.042 (0.004)	0.787 (0.005)
Linear Reg. (Trunc.)	0.192 (0.002)	0	0	0.808 (0.002)
Linear Reg. (Shrink)	0.192 (0.002)	0.034 (0.003)	0.030 (0.003)	0.804 (0.004)
Convex Surrogate	0.193 (0.002)	0.043 (0.003)	0.026 (0.003)	0.794 (0.004)
BFGS	0.193 (0.002)	0.035 (0.003)	0.021 (0.003)	0.796 (0.004)
Linear Reg. (Balanced)	0.192 (0.002)	0.048 (0.004)	0.030 (0.003)	0.793 (0.005)
Linear Reg. (Group)	0.192 (0.002)	0.045 (0.004)	0.033 (0.003)	0.798 (0.005)
XG Boost	0.192 (0.002)	0.045 (0.004)	0.030 (0.003)	0.796 (0.005)
XG Boost (Trunc.)	0.193 (0.002)	0	0	0.807 (0.002)
XG Boost (Shrink)	0.192 (0.002)	0.033 (0.003)	0.030 (0.003)	0.804 (0.003)

Table 5: Mean square error (MSE), false positive rates (FP) by race, and NIJFM scores on held-out (50%) test data with competition cutoff of 0.5 for decision boundary. Bootstrap standard errors reported in parentheses.

5.2 Did the 0.5 threshold affect anything? Would your team recommend a different threshold?

In Table 5 we display results for several different models on data for the NIJ competition using a decision threshold of 0.5. Full details on the models can be found in [10]. We include the MSE, false postive rates (FP), and the NIJFM scores along with bootstrap standard errors[2]. Differences in the MSE across models are not statistically significant, with all models achieving a held-out MSE of 0.192-0.193. NIJFM scores range from 0.787-0.808, with simple truncation or shrinkage applied to linear regression and xgboost having as good or better fairness scores compared to the other approaches. As noted above, in the competition we used truncation on the female scores and optimized MSE with no fairness adjustment for male scores.

We note that the false positive rates in Table 5 are low across all methods. This is due to the fact that the decision cutoff of 0.5 is far from the base rate of recidivism for the dataset (0.298). To investigate the sensitivity of results to the decision cutoff further, in Table 6 we display results for different models on held-out test data for cutoffs of 0.3 (corresponding to more false positives and less false negatives) and 0.7 (corresponding to less false positives and more false negatives). As the cutoff moves further away from the base rate of recidivism, we see less of a difference in NIJFM scores across fairness-aware regressions and the standard linear/logistic regressions. This is because fewer individuals are forecasted to be above the decision boundary, and therefore the false positive rates are much lower (and approach zero as the decision boundary moves further from the base rate).

The threshold in practice will depend on the particular application. However, in the competition a threshold closer to the base rate of 0.3 would have led to more emphasis placed on the fairness penalty

[2] Bootstrap standard errors are calculated for each model fit to the training data by resampling the test data with replacement 1000 times and calculating the standard deviation of the statistic across samples.

Model	FP Black ($c = 0.3$)	FP white ($c = 0.3$)	NIJFM ($c = 0.3$)	FP Black ($c = 0.7$)	FP white ($c = 0.7$)	NIJFM ($c = 0.7$)
Linear Reg.	0.485 (0.009)	0.354 (0.009)	0.702 (0.010)	0	0	0.808 (0.002)
Logistic Reg.	0.430 (0.008)	0.303 (0.009)	0.706 (0.010)	0.003 (0.001)	0.001 (0.001)	0.807 (0.002)
Linear Reg. (Trunc.)	0	0	0.798 (0.002)	0	0	0.808 (0.002)
Linear Reg. (Shrink)	0.369 (0.008)	0.354 (0.009)	0.795 (0.008)	0	0	0.808 (0.002)
Convex Surrogate	0.478 (0.009)	0.377 (0.009)	0.725 (0.010)	0	0	0.808 (0.002)
BFGS	0.434 (0.008)	0.428 (0.010)	0.795 (0.007)	0	0	0.808 (0.002)
Linear Reg. (Balanced)	0.487 (0.008)	0.354 (0.009)	0.701 (0.010)	0	0	0.808 (0.002)
Linear Reg. (Group)	0.485 (0.008)	0.358 (0.009)	0.705 (0.010)	0	0	0.808 (0.002)
XG Boost	0.472 (0.008)	0.377 (0.009)	0.731 (0.010)	0	0	0.808 (0.002)
XG Boost (Trunc.)	0	0	0.798 (0.002)	0	0	0.808 (0.002)
XG Boost (Shrink)	0.386 (0.008)	0.377 (0.009)	0.798 (0.007)	0	0	0.808 (0.002)

Table 6: NIJFM scores and false positive rates on held-out (50%) test data with cutoffs of $c = 0.3$ and $c = 0.7$ for decision boundary. Bootstrap standard errors reported in parentheses.

component of the NIJFM. This possibly would have led to more innovations in fairness-encouraging recidivism forecasting.

5.3 Did the fact that the fairness penalty only considered false positives affect your submission?

The fact that the fairness penalty only considered false positives affected our submission by encouraging the use of truncation to minimize the FPR. In particular, for the female category there were so few individuals above the 0.5 cutoff that the FPR had high variance and it was advantageous to set those scores to 0.4999. If, for example, false negative rates were used instead, then at a cutoff of 0.5 the FNR would have contributed much more to the penalty since so many individuals are below the cutoff (and only those below will contribute to false negatives). In that case we would not have used truncation above the threshold and would have selected a different fairness-encouraging technique.

6 Conclusion

6.1 Are there practical/applied findings that could help the field based on your work? If yes, what are they?

Event level data that was available after parole seemed to be stronger features than static demographic data. So in practice generating a good feature set will be important for buidling accurate forecasts. This is echoed in recent research, where humans can outperform models with limited features, however algorithms outperform humans when the feature set is expanded [8].

6.2 What should NIJ have considered changing (other than metrics) to improve this Challenge?

In addition to expanding the set of features, NIJ could have considered changing the task from binary classification to a survival modeling set up where one predicts also the time to recidivism. Again, having event level data where each event has a time and a feature could be useful for building such models. This also would have made a contribution to research on algorithmic fairness, as research on fairness-aware survival modeling and time-to-event prediction is less developed than binary classification.

6.3 For future Challenges, what should NIJ consider changing to improve Challenges? For example, more/less time, different topic, or data issues (missing data)?

In the future NIJ might consider a different topic. A timely focus of a competition could be forecasting officer excessive use of force and/or misconduct.

References

[1] Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

[2] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning. (FATML) 2017.*, 2017.

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.

[4] Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

[5] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[6] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[7] Edward W Frees, Glenn Meyers, and A David Cummings. Summarizing insurance scores using a gini index. *Journal of the American Statistical Association*, 106(495):1085–1098, 2011.

[8] Jongbin Jung, Sharad Goel, Jennifer Skeem, et al. The limits of human predictions of recidivism. *Science Advances*, 6(7):eaaz0652, 2020.

[9] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[10] George Mohler and Michael D Porter. A note on the multiplicative fairness score in the nij recidivism forecasting challenge. *Crime Science*, 10(1):1–5, 2021.

[11] Robert Tibshirani and Keith Knight. Model search by bootstrap "bumping". *Journal of Computational and Graphical Statistics*, 8(4):671–686, 1999.

[12] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio Calmon. Optimized score transformation for fair classification. In *International Conference on Artificial Intelligence and Statistics*, pages 1673–1683. PMLR, 2020.