Supplementary Materials

for

Can non-experts really emulate statistical learning methods? A comment on “The accuracy, fairness, and limits of predicting recidivism”

Kirk Bansak

Additional Technical Details

Data

This study exclusively employs data used in Dressel and Farid (2018). This includes a data set of 7214 pretrial criminal defendants in Broward County, Florida from 2013 to 2014, which contains the defendants' demographic information, criminal history, and whether or not they recidivated within the following two years. In addition, it includes data from surveys of Amazon Mechanical Turkers (MTurkers) who were recruited to evaluate 1000 randomly sampled defendants from the Broward County data set. Further details and all data can be accessed via the documentation provided in Dressel and Farid (2018).

Boosted Trees Model Fitting

In this study, the stochastic gradient boosted trees models were implemented using the `gbm` package in `R`, a bag fraction of 0.5, and binomial deviance loss function. 6-fold cross-validation was employed within the training data to select tuning parameter values, including the interaction depth (interaction depths of 3, 4, and 5 were tested) and number of boosting iterations (early stopping point). The learning rate (shrinkage) was held fixed at 0.01.

Predictors

The predictors employed for both the logistic regression and boosted trees models are age, sex, number of juvenile misdemeanors, number of juvenile felonies, number of prior (nonjuvenile) crimes, crime degree (misdemeanor vs. felony), and crime charge. In the data used in the original study, crime charge is an unordered categorical variable (146 distinct summary labels are employed). The simplest way to handle an unordered categorical variable of this sort is to transform it into a set of dummy indicators for each category. However, this is not always good practice for high-cardinality categorical variables with sparseness in many of the categories, as it not only increases computational costs but is also known to pose the risk of over-fitting and model instability. Accordingly, various automated methods of processing such variables as part of or prior to model fitting—via ordering, scaling, clustering/fusion, etc.—have been proposed and employed in the existing literature (Breiman et al., 1984; Quinlan, 1986; Tibshirani et al., 2005; Gertheiss et al., 2010; Micci-Barreca, 2001; Bondell and Reich, 2009).

Given that crime charge is both high-cardinality and sparse in many categories (96 of the 146 categories have 10 or fewer observations, and 48 categories have only 1 observation), one such automated processing method is employed here. That is, each category is mapped to a scalar value between 0 and 1, specifically a posterior estimate of the proportion in the positive outcome class using a beta-Bernoulli empirical Bayes model (Micci-Barreca, 2001). Importantly, the mapping is determined exclusively using training data, with both the empirical prior and proportion of positivies within categories determined only using training data, and then applied to test data; that is, test data are never observed in the creation of the mapping rules. Given the binary partitioning process employed by decision trees, this strategy is similar to a commonly used approach (Breiman et al., 1984; Friedman et al., 2009) of determining unordered categorical variable splits by converting them into ordered variables according to the proportion in each category falling into the positive outcome class. The empirical Bayes approach used here has an advantage in the case of sparse categories, as it allows for shrinkage to the empirical prior (the full training data proportion in the positive outcome class) and also naturally handles cases in which a category that appears in the test data does not appear in the training data (i.e. the empirical Bayes approach will simply treat this as a case of zero observations in the category in question and hence put all weight on the prior).

In addition, this Supplementary Materials (SM) document also includes results (Table S.1, Figures S.3 and S.4) for when crime charge is simply converted into dummy indicators. Those results are roughly similar though a bit worse, suggesting the use of dummy indicators indeed leads to overfitting here, particularly for the logistic regression.

Additional Details on Modeling Uncertainty

95% confidence intervals for all results of the statistical learning methods are constructed using the empirical percentiles (2.5 and 97.5) across the 1000 evaluations.

For the *Sample* approach to modeling uncertainty, described in the main text, tuning parameters for the boosted trees models were chosen via cross-validation on the training data separately for each evaluation.

For the *Bootstrap* approach to modeling uncertainty, described in the main text, the same tuning parameter values were used for all evaluations of the boosted trees models and were chosen via cross-validation on the intact training data set (i.e. the original data without the 1000 test units evaluated by the MTurkers). The reason that tuning parameters were not chosen via cross-validation separately for each evaluation under this approach is that the desirable statistical properties of cross-validation depend upon the cross-validation subsets being non-overlapping, whereas this would not be the case in a bootstrapped re-sample, in which over one third of the re-sampled units are duplicates on average. Performing cross-validation on such a re-sample would lead to overfitting.

Log Score Calculations

The logarithmic scoring rule is $\mathbf{I}(y_i = 1)ln(p_i) + \mathbf{I}(y_i = 0)ln(1 - p_i)$, and the statistical results reported are for the mean logarithmic score. For the MTurkers' evaluations, the mean score is unfortunately undefined (approaching $-\infty$ in the limit) given the existence of predicted probabilities that equal 0 (1) for units that do (do not) recidivate. The results reported for the MTurkers are the mean logarithmic score omitting such units, thus providing for a harder test that is biased in favor of the MTurkers' evaluations.

Replication Archive

Replication materials are available in Bansak (2018).

Additional Results

The following tables and figures present the results of additional analyses referred to in the main text and in this SM.

Table S.1: **Model performance results, using dummy indicators for crime charge.** The table displays several performance metrics for the statistical learning methods—gradient boosted trees (GBM) and logistic regression (Logit)—under both approaches to modeling uncertainty (*Sample* and *Bootstrap*), along with the results for the MTurkers' pooled evaluations both without and with race presented. For the statistical learning methods, 95% confidence intervals are displayed. A cut point of 0.5 is employed for the PCC, FPR, and FNR.

Statistical Method	Uncertainty Method	PCC	AUC-ROC	FPR	FNR	Brier Score	Log Score
Logit	Sample	$[0.634, 0.693]$	$[0.642, 0.743]$	$[0.181, 0.270]$	$[0.411, 0.562]$	$[0.207, 0.261]$	$[-1.744, -0.653]$
GBM[a]	Sample	$[0.660, 0.713]$	$[0.708, 0.767]$	$[0.205, 0.282]$	$[0.355, 0.443]$	$[0.195, 0.216]$	$[-0.621, -0.576]$
Logit	Bootstrap	$[0.658, 0.687]$	$[0.711, 0.734]$	$[0.208, 0.288]$	$[0.372, 0.462]$	$[0.212, 0.222]$	$[-0.868, -0.723]$
GBM	Bootstrap	$[0.676, 0.699]$	$[0.741, 0.752]$	$[0.206, 0.271]$	$[0.366, 0.426]$	$[0.202, 0.206]$	$[-0.601, -0.591]$
MTurk (w/o race)	–	0.670	0.709	0.323	0.338	0.240	-0.669
MTurk (w/ race)	–	0.665	0.709	0.324	0.347	0.240	-0.658

[a] To lower computational costs, which increase substantially with the conversion of the crime charge categorical variable into dummy indicators, the number of cross-validation folds is decreased to 3 and the interaction depth is held fixed at 5 for the implementation of the *Sample* approach to modeling uncertainty with the gradient boosted trees model.

Table S.2: **Model performance results, specifying binary classification criterion (cut point) to balance false positive and negative rates.** The table displays several performance metrics for the statistical learning methods—gradient boosted trees (GBM) and logistic regression (Logit)—under both approaches to modeling uncertainty (*Sample* and *Bootstrap*), using classification criteria (cut points) chosen with a precision of three significant digits to balance the mean values of the false positive rate (FPR) and false negative rate (FNR). Mean values and 95% confidence intervals are displayed.

Statistical Method	Uncertainty Method	Cut Point	PCC	FPR	FNR
Logit	Sample	0.451	0.663	0.337	0.336
			$[0.636, 0.693]$	$[0.293, 0.377]$	$[0.293, 0.382]$
GBM	Sample	0.429	0.675	0.325	0.324
			$[0.648, 0.701]$	$[0.287, 0.363]$	$[0.281, 0.368]$
Logit	Bootstrap	0.440	0.664	0.337	0.336
			$[0.653, 0.675]$	$[0.298, 0.374]$	$[0.307, 0.359]$
GBM	Bootstrap	0.434	0.677	0.323	0.323
			$[0.665, 0.689]$	$[0.292, 0.355]$	$[0.296, 0.351]$

Results for the AUC-ROC, Brier Score, and Log Score are not displayed as those metrics are not computed as a function of a binary classification criterion and hence remain unchanged from Table 1 in the main text.

Table S.3: **Model performance results, for black defendants only.** The table displays several performance metrics for the statistical learning methods—gradient boosted trees (GBM) and logistic regression (Logit)—under both approaches to modeling uncertainty (*Sample* and *Bootstrap*), along with the results for the MTurkers' pooled evaluations both without and with race presented, as applied to black defendants only. For the statistical learning methods, 95% confidence intervals are displayed. A cut point of 0.5 is employed for the PCC, FPR, and FNR.

Statistical Method	Uncertainty Method	PCC	AUC-ROC	FPR	FNR	Brier Score	Log Score
Logit	Sample	$[0.635, 0.711]$	$[0.681, 0.763]$	$[0.244, 0.358]$	$[0.288, 0.407]$	$[0.204, 0.228]$	$[-0.653, -0.593]$
GBM	Sample	$[0.641, 0.718]$	$[0.695, 0.773]$	$[0.271, 0.386]$	$[0.259, 0.363]$	$[0.195, 0.223]$	$[-0.636, -0.575]$
Logit	Bootstrap	$[0.658, 0.692]$	$[0.714, 0.728]$	$[0.202, 0.311]$	$[0.334, 0.430]$	$[0.216, 0.220]$	$[-0.634, -0.626]$
GBM	Bootstrap	$[0.653, 0.687]$	$[0.708, 0.734]$	$[0.289, 0.373]$	$[0.298, 0.361]$	$[0.209, 0.219]$	$[-0.632, -0.607]$
MTurk (w/o race)	–	0.679	0.699	0.360	0.291	0.240	-0.670
MTurk (w/ race)	–	0.658	0.689	0.399	0.298	0.247	-0.665

Table S.4: **Model performance results, for white defendants only.** The table displays several performance metrics for the statistical learning methods—gradient boosted trees (GBM) and logistic regression (Logit)—under both approaches to modeling uncertainty (*Sample* and *Bootstrap*), along with the results for the MTurkers' pooled evaluations both without and with race presented, as applied to white defendants only. For the statistical learning methods, 95% confidence intervals are displayed. A cut point of 0.5 is employed for the PCC, FPR, and FNR.

Statistical Method	Uncertainty Method	PCC	AUC-ROC	FPR	FNR	Brier Score	Log Score
Logit	Sample	$[0.632, 0.721]$	$[0.645, 0.753]$	$[0.102, 0.194]$	$[0.516, 0.673]$	$[0.194, 0.228]$	$[-0.652, -0.571]$
GBM	Sample	$[0.636, 0.728]$	$[0.653, 0.764]$	$[0.129, 0.230]$	$[0.459, 0.606]$	$[0.188, 0.227]$	$[-0.648, -0.560]$
Logit	Bootstrap	$[0.682, 0.706]$	$[0.721, 0.735]$	$[0.118, 0.177]$	$[0.529, 0.650]$	$[0.197, 0.201]$	$[-0.587, -0.580]$
GBM	Bootstrap	$[0.684, 0.721]$	$[0.720, 0.747]$	$[0.160, 0.224]$	$[0.436, 0.521]$	$[0.193, 0.201]$	$[-0.592, -0.570]$
MTurk (w/o race)	–	0.671	0.705	0.274	0.421	0.234	-0.653
MTurk (w/ race)	–	0.674	0.708	0.262	0.436	0.226	-0.635

Table S.5: **Model performance results, specifying binary classification criterion (cut point) to balance false positive and negative rates, for black defendants only.** The table displays several performance metrics for the statistical learning methods—gradient boosted trees (GBM) and logistic regression (Logit)—under both approaches to modeling uncertainty (*Sample* and *Bootstrap*), using classification criteria (cut points) chosen with a precision of three significant digits to balance the mean values of the false positive rate (FPR) and false negative rate (FNR), as applied to black defendants only. Mean values and 95% confidence intervals are displayed.

Statistical Method	Uncertainty Method	Cut Point	PCC	FPR	FNR
Logit	Sample	0.492	0.673 [0.635, 0.711]	0.326 [0.268, 0.381]	0.328 [0.271, 0.382]
GBM	Sample	0.508	0.681 [0.641, 0.717]	0.319 [0.263, 0.378]	0.319 [0.266, 0.373]
Logit	Bootstrap	0.475	0.676 [0.660, 0.691]	0.323 [0.263, 0.386]	0.324 [0.295, 0.358]
GBM	Bootstrap	0.501	0.670 [0.653, 0.689]	0.330 [0.285, 0.373]	0.330 [0.298, 0.361]

Results for the AUC-ROC, Brier Score, and Log Score are not displayed as those metrics are not computed as a function of a binary classification criterion and hence remain unchanged from Table S.3.

Table S.6: **Model performance results, specifying binary classification criterion (cut point) to balance false positive and negative rates, for white defendants only.** The table displays several performance metrics for the statistical learning methods—gradient boosted trees (GBM) and logistic regression (Logit)—under both approaches to modeling uncertainty (*Sample* and *Bootstrap*), using classification criteria (cut points) chosen with a precision of three significant digits to balance the mean values of the false positive rate (FPR) and false negative rate (FNR), as applied to white defendants only. Mean values and 95% confidence intervals are displayed.

Statistical Method	Uncertainty Method	Cut Point	PCC	FPR	FNR
Logit	Sample	0.400	0.642 $[0.596, 0.689]$	0.358 $[0.296, 0.422]$	0.359 $[0.285, 0.436]$
GBM	Sample	0.360	0.658 $[0.609, 0.707]$	0.342 $[0.277, 0.406]$	0.343 $[0.264, 0.423]$
Logit	Bootstrap	0.389	0.675 $[0.660, 0.692]$	0.326 $[0.295, 0.363]$	0.324 $[0.286, 0.364]$
GBM	Bootstrap	0.370	0.684 $[0.660, 0.706]$	0.316 $[0.278, 0.359]$	0.316 $[0.271, 0.357]$

Results for the AUC-ROC, Brier Score, and Log Score are not displayed as those metrics are not computed as a function of a binary classification criterion and hence remain unchanged from Table S.4.

Figure S.1: **Probability calibration across methods (MTurkers told defendant race), using *Sample* approach to model uncertainty.** The top two panels display probability calibration plots. Each point and interval in the upper two panels correspond to a bin of predicted probabilities. The black triangles comprise the MTurkers' calibration points for the evaluations where MTurkers were provided with the defendants' race. Each point's position along the x-axis signifies the mean predicted probability within the bin, while its position on the y-axis signifies the actual proportion of positives among the units contained within the bin. The gray points represent the mean proportion of positives within each bin across 1000 evaluations of each of the statistical learning methods, while the error bars provide 95% confidence intervals for the proportion of positives within each bin, with uncertainty modeled using the *Sample* approach described in the main text. The three bottom panels display histograms of the predicted probabilities for each method.

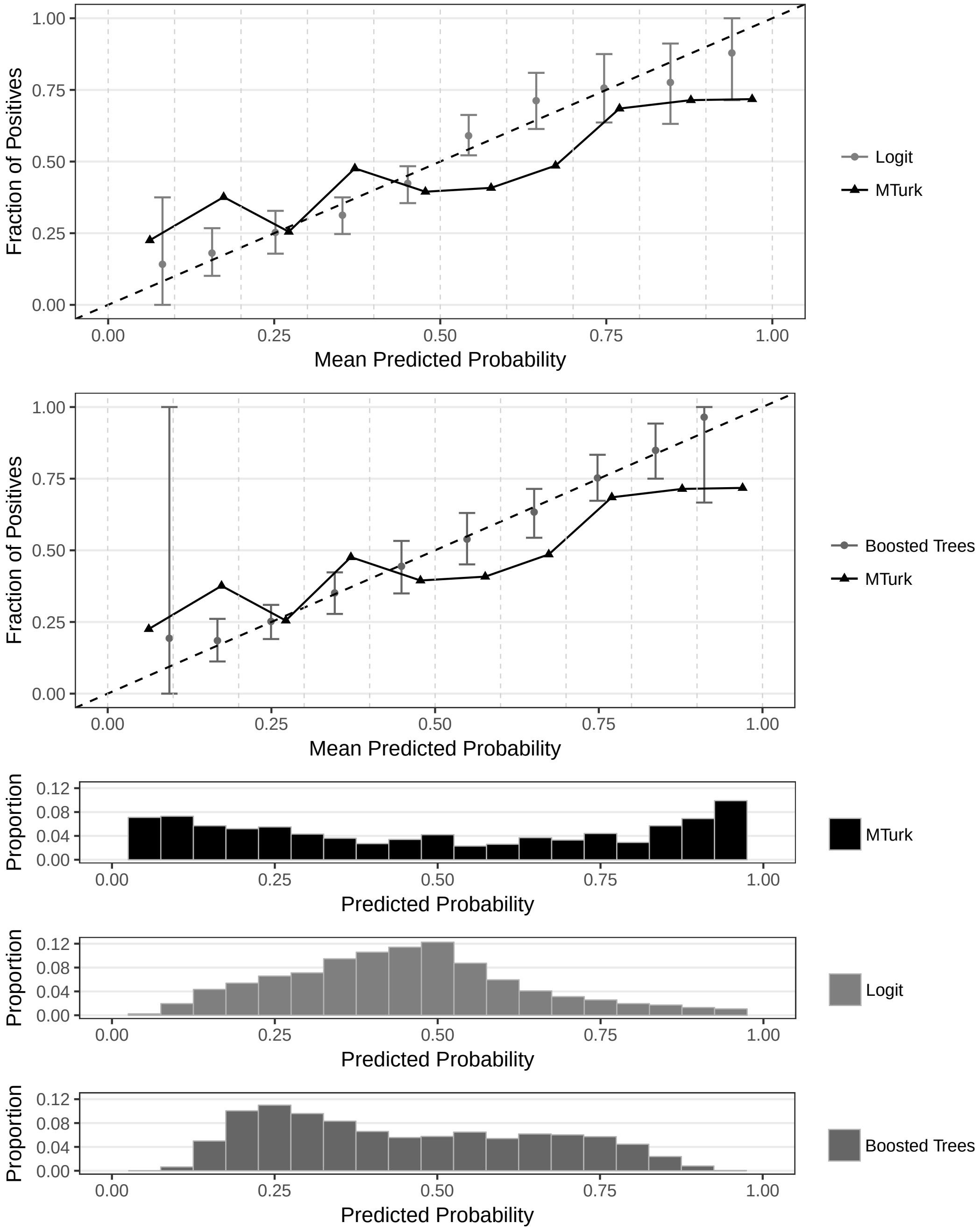


Figure S.2: **Probability calibration across methods (MTurkers told defendant race), using *Bootstrap* approach to model uncertainty.** The top two panels display probability calibration plots. Each point and interval in the upper two panels correspond to a bin of predicted probabilities. The black triangles comprise the MTurkers' calibration points for the evaluations where MTurkers were provided with the defendants' race. Each point's position along the x-axis signifies the mean predicted probability within the bin, while its position on the y-axis signifies the actual proportion of positives among the units contained within the bin. The gray points represent the mean proportion of positives within each bin across 1000 evaluations of each of the statistical learning methods, while the error bars provide 95% confidence intervals for the proportion of positives within each bin, with uncertainty modeled using the *Bootstrap* approach described in the main text. The three bottom panels display histograms of the predicted probabilities for each method.

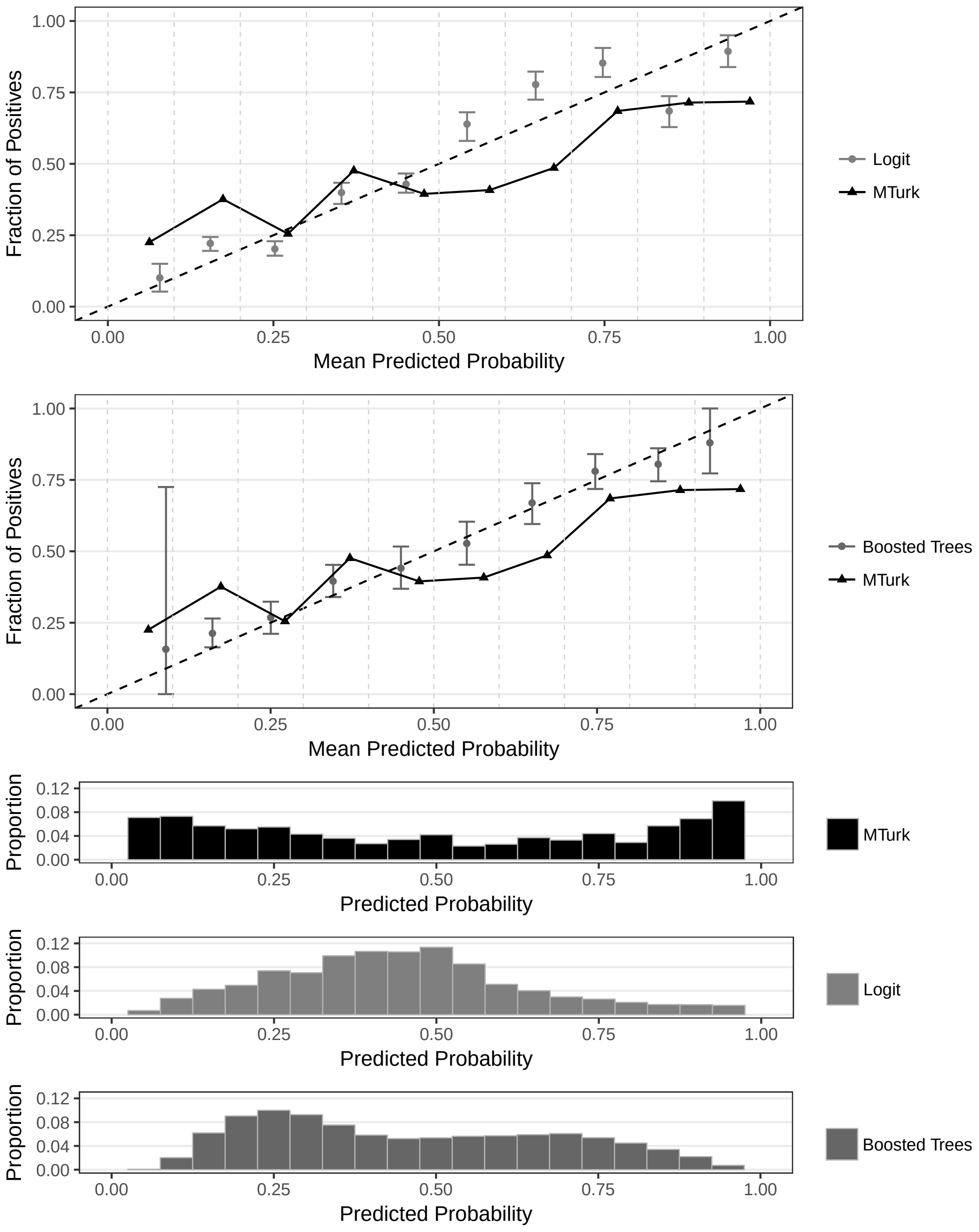


Figure S.3: **Probability calibration across methods (MTurkers not told defendant race), using *Sample* approach to model uncertainty and dummy indicators for crime charge.** The top two panels display probability calibration plots. Each point and interval in the upper two panels correspond to a bin of predicted probabilities. The black triangles comprise the MTurkers' calibration points for the evaluations where MTurkers were not provided with the defendants' race. Each point's position along the x-axis signifies the mean predicted probability within the bin, while its position on the y-axis signifies the actual proportion of positives among the units contained within the bin. The gray points represent the mean proportion of positives within each bin across 1000 evaluations of each of the statistical learning methods, while the error bars provide 95% confidence intervals for the proportion of positives within each bin, with uncertainty modeled using the *Sample* approach described in the main text. The three bottom panels display histograms of the predicted probabilities for each method.

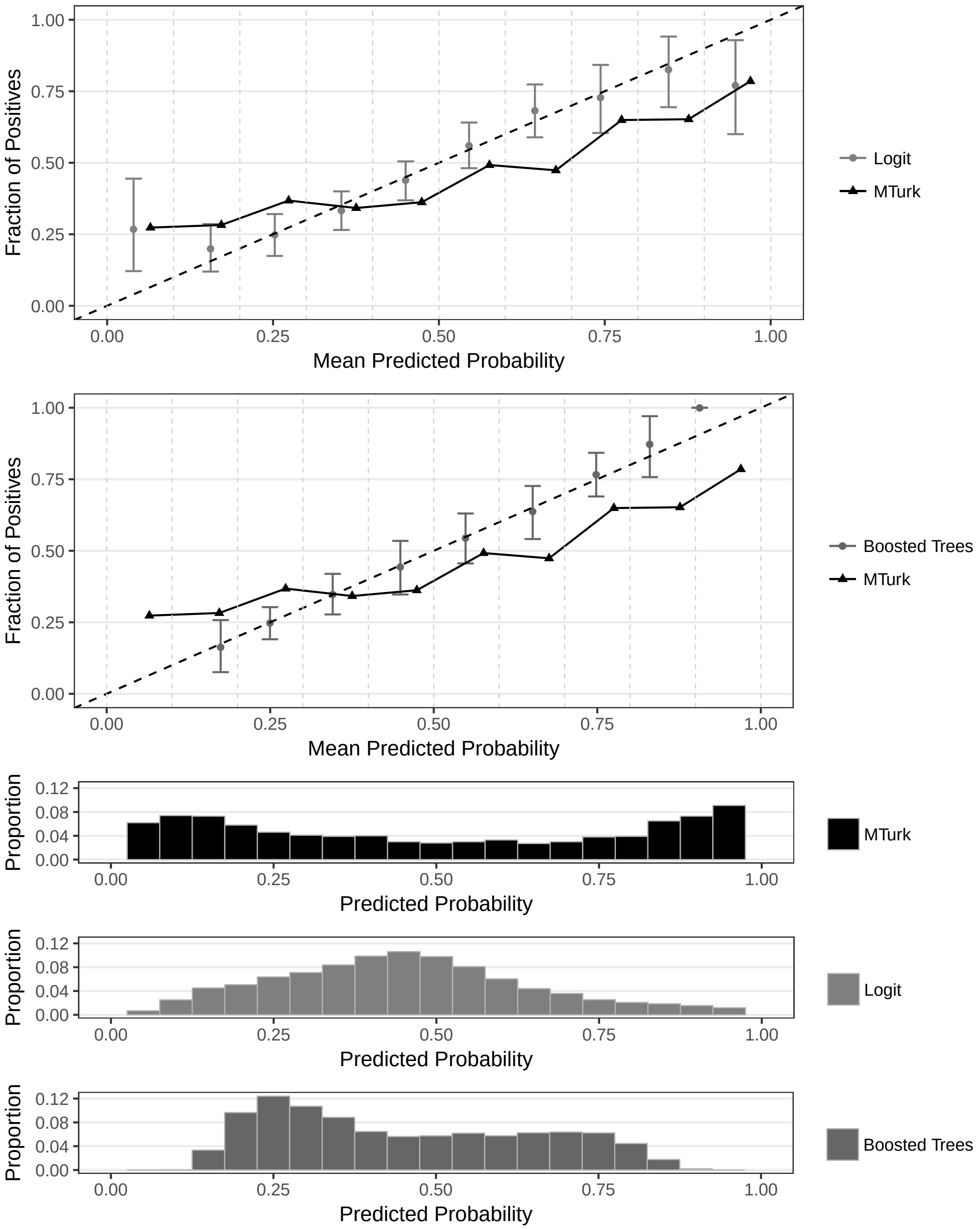
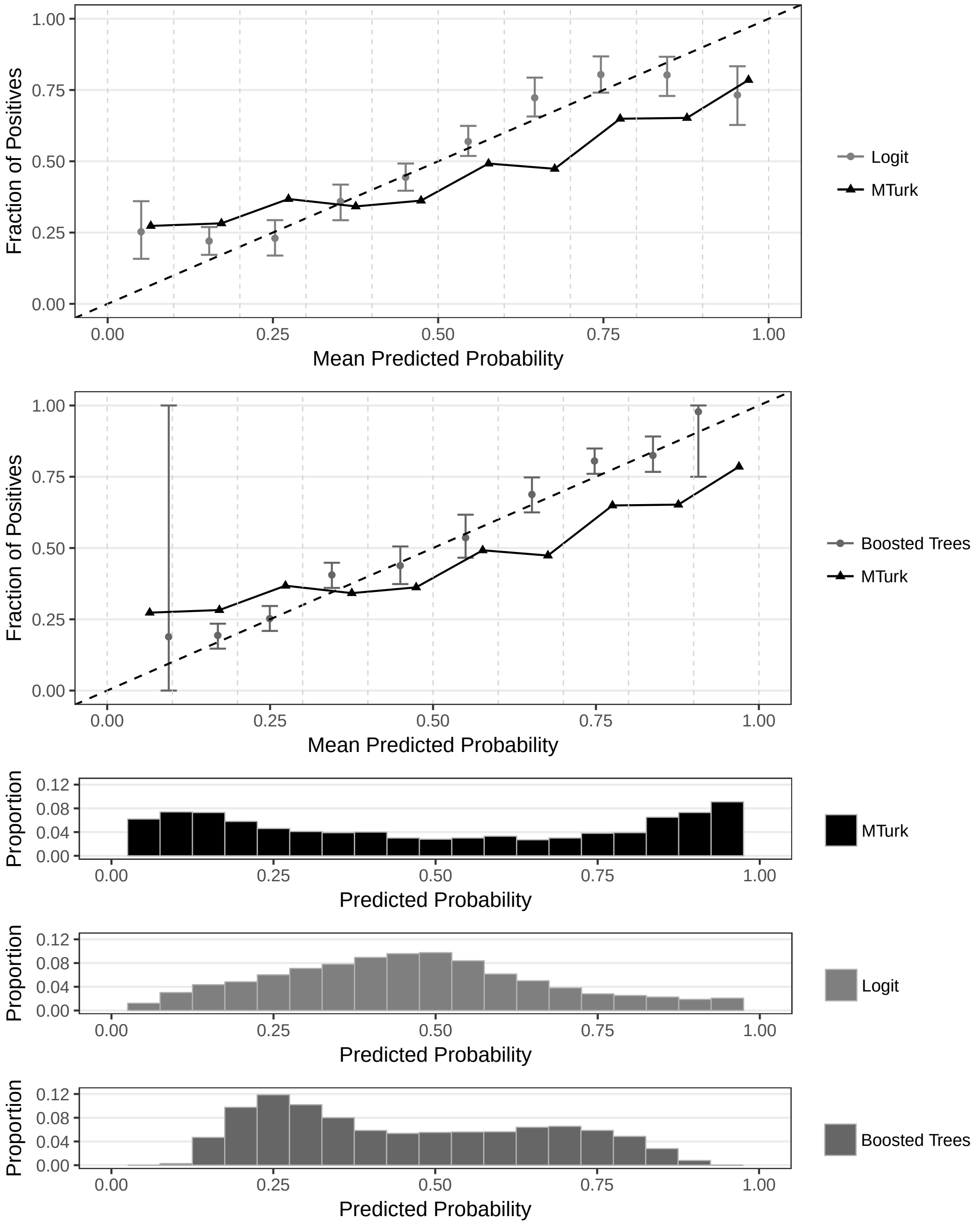


Figure S.4: **Probability calibration across methods (MTurkers not told defendant race), using *Bootstrap* approach to model uncertainty and dummy indicators for crime charge.** The top two panels display probability calibration plots. Each point and interval in the upper two panels correspond to a bin of predicted probabilities. The black triangles comprise the MTurkers' calibration points for the evaluations where MTurkers were not provided with the defendants' race. Each point's position along the x-axis signifies the mean predicted probability within the bin, while its position on the y-axis signifies the actual proportion of positives among the units contained within the bin. The gray points represent the mean proportion of positives within each bin across 1000 evaluations of each of the statistical learning methods, while the error bars provide 95% confidence intervals for the proportion of positives within each bin, with uncertainty modeled using the *Bootstrap* approach described in the main text. The three bottom panels display histograms of the predicted probabilities for each method.



References

Bansak, Kirk (2018). Replication materials for: Can non-experts really emulate statistical learning methods? *Harvard Dataverse*. doi: 10.7910/DVN/KT20FE.

Bondell, Howard D. and Brian J. Reich (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics 65*(1), 169–177.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone (1984). *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth.

Dressel, Julia and Hany Farid (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances 4*(1), eaao5580.

Friedman, Jerome H., Trevor Hastie, and Robert Tibshirani (2009). *The Elements of Statistical Learning, 2nd ed.* Springer.

Gertheiss, Jan, Gerhard Tutz, et al. (2010). Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics 4*(4), 2150–2180.

Micci-Barreca, Daniele (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter 3*(1), 27–32.

Quinlan, J. Ross (1986). Induction of decision trees. *Machine Learning 1*(1), 81–106.

Tibshirani, Robert, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(1), 91–108.