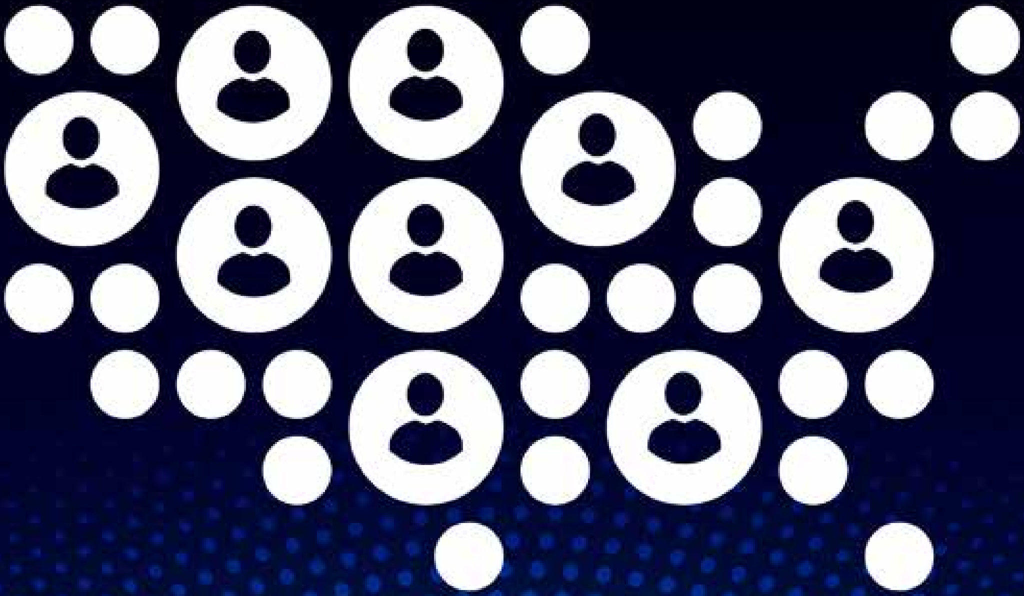
U.S. Department of Justice
Office of Justice Programs
National Institute of Justice



NATIONAL INSTITUTE OF JUSTICE

THE NIJ RECIDIVISM FORECASTING CHALLENGE: CONTEXTUALIZING THE RESULTS

Veronica White
D. Michael Applegarth
Joel Hunt
Caleb Hudgins

February 2022

NIJ.OJP.GOV | National Institute of Justice
STRENGTHEN SCIENCE. ADVANCE JUSTICE.

U.S. Department of Justice
Office of Justice Programs
810 Seventh St. N.W.
Washington, DC 20531

Jennifer Scherer, Ph.D.
Acting Director, National Institute of Justice

This and other publications and products of the National Institute of Justice can be found at:

National Institute of Justice
Strengthen Science • Advance Justice
NIJ.ojp.gov

Office of Justice Programs
Building Solutions • Supporting Communities • Advancing Justice
OJP.gov

The National Institute of Justice is the research, development, and evaluation agency of the U.S. Department of Justice. NIJ's mission is to advance scientific research, development, and evaluation to enhance the administration of justice and public safety.

The National Institute of Justice is a component of the Office of Justice Programs, which also includes the Bureau of Justice Assistance; the Bureau of Justice Statistics; the Office for Victims of Crime; the Office of Juvenile Justice and Delinquency Prevention; and the Office of Sex Offender Sentencing, Monitoring, Apprehending, Registering, and Tracking.

Opinions or conclusions expressed in this paper are those of the authors and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Contextualizing the Fiscal Year 2021 NIJ Recidivism Forecasting Challenge Results

Abstract

The National Institute of Justice (NIJ) recently hosted the Fiscal Year 2021 Recidivism Forecasting Challenge. The primary aim of this research competition was to increase public safety and the fair administration of justice by improving the ability to forecast and understand the variables that impact the likelihood that an individual under parole supervision will recidivate. Entrants were provided with two datasets. The first was a training dataset of over 18,000 individuals released from prison to parole supervision in the state of Georgia during the period of January 1, 2013, through December 31, 2015. These data contained information about individuals' demographic characteristics, supervision case information, prison case information, prior criminal and community supervision history in the state of Georgia, activities for current supervision, and whether they recidivated in any of the three years after they began supervision.

The second was a test dataset (n = 7,807) used to develop models for forecasting the probability that an individual on parole will recidivate within their first, second, or third year on parole. For each of the Challenge's three submission periods, models were scored by two indices: (1) a Brier score, which is a measure of accuracy, and (2) fairness and accuracy via a difference in the false positive rate between Black and white racial groups in conjunction with the Brier score. Prizes were awarded to the entries that had the lowest error in the forecasts for males and females, and the average of these two scores. Additionally, prizes were awarded to the entries that had the highest fairness and accuracy scores after any assessed fairness penalties.

In order to put these results into context, this paper compares the winning results to a variety of naive models, such as predicting recidivism by random chance or using the average recidivism rate by population demographic for those in the sample. Naive demographic models outperformed the chance model, and submitted forecasts outperformed the best naive demographic models. This suggests that more advanced algorithms have improved the capability of determining which variables accurately forecast recidivism. Improved algorithms could assist community corrections agencies in identifying and prioritizing the needs of those on parole and promoting more successful reintegration into society. Future papers will consider alternative metrics for fairness and accuracy, provide a more detailed comparison of submissions, and explore practical implications for predicting recidivism.

Introduction

The National Institute of Justice (NIJ) recently hosted the Fiscal Year 2021 Recidivism Forecasting Challenge. The primary aim of the Challenge was to increase public safety and the fair administration of justice by understanding the factors that drive recidivism. To do so, Challenge entrants were given a dataset that allowed them to explore gender, racial, and age differences for individuals on parole, in addition to a host of other variables described below. The Challenge was designed to improve the accuracy and fairness of forecasts by identifying key variables, accounting for gender-specific needs, and adjusting for potential racial bias in predicting recidivism. NIJ's aim was to increase public safety and the fair administration of justice by improving the ability to forecast and understand the variables that impact the likelihood that an individual under parole supervision will recidivate. The Challenge encouraged data scientists from all fields to build upon the current knowledge base for forecasting recidivism while also infusing innovative methods and new perspectives. Here, NIJ provides a brief overview of the Challenge and the metrics used to judge the entries, and contextualizes how the winners' forecasts performed compared to several naive models in terms of accuracy and fairness. A detailed description of the Challenge is available on the NIJ website.[1]

[1] "Recidivism Forecasting Challenge," *National Institute of Justice*, https://nij.ojp.gov/funding/recidivism-forecasting-challenge.

Challenge Background

The Challenge used data from the state of Georgia that contained records for over 25,000 persons released from prison to parole supervision during the period of January 1, 2013, through December 31, 2015.[2] The dataset included the following variable categories: supervision case information (e.g., gender, race, and age group), prison case information (e.g., education level upon entry, crime of conviction), prior Georgia criminal history (e.g., number of prior arrests, type of arrests), prior Georgia community supervision history (e.g., prior revocations), conditions of current supervision period (e.g., mental health or substance abuse programming), supervision activities for current supervision period (e.g., number of delinquency reports, total number of all program attendances), and four measures of recidivism across three years of supervision. Recidivism was defined as an arrest for a new felony or misdemeanor charge once the supervision period started.[3] Not all individuals in the dataset were under supervision for the entire three years of the dataset; however, if individuals were arrested for new felony or misdemeanor crimes within three years of the start of their parole, they were still classified as recidivating. Geographic information was also provided pertaining to where individuals were initially released on parole. In order to minimize the risk of disclosing personally identifiable information, a combined public use microdata area[4] was provided for each individual. Merging additional outside data sources into the Challenge's dataset was encouraged, allowing each submission to incorporate unique variables.

[2] The Challenge data only contain individuals with the racial categories of Black and white. In the original dataset, fewer than 500 individuals were identified as Hispanic, and fewer than 100 individuals each were identified as Asian, Native American, other, or unknown. All five of these categories were dropped from the sample to prevent inadvertent disclosure of personal identifying information. It should also be noted that the state identified individuals' race, ethnicity, and gender, meaning individuals may not have self-identified with the categories in which they were ultimately classified. Among other things, this potentially impacted the number of individuals identified as Hispanic, as they may have been labeled as white. NIJ maintained the naming conventions for the variables provided by the state of Georgia (e.g., gender, race). For more information on how the Challenge dataset was prepared, visit https://nij.ojp.gov/funding/recidivism-forecasting-challenge#appendices.

[3] For a complete description of the variables contained in the dataset, see the Challenge codebook at https://nij.ojp.gov/funding/recidivism-forecasting-challenge#recidivism-forecasting-challenge-database-fields-defined.

[4] A public use microdata area (PUMA) is based on aggregations of counties and census tracts. Each PUMA must have a population of 100,000 or more throughout the census decade, and the building blocks for the PUMA must be continuous, unless the area is an island. For this Challenge, PUMA blocks were collapsed to reduce risk of identification. For more information on how this was done, visit https://nij.ojp.gov/funding/recidivism-forecasting-challenge#j55a9e. For more information on how PUMAs are developed and used, visit https://www.census.gov/programs-surveys/acs/guidance/handbooks/pums.html.

Descriptive Statistics of Provided Datasets

For the Challenge, the dataset was split into two subsets: training and test data. NIJ used a random 70/30 split, placing 70% of the total dataset in the training set and 30% in the test set. NIJ ensured that the random split produced equal distributions on key variables (e.g., recidivism, gender, age, race, education, prerelease risk scores). NIJ also conducted a disclosure risk analysis to ensure appropriate steps were taken to reduce the risk of deductive disclosure of individuals' identities.[5] Exhibits 1 and 2 provide the descriptive statistics for the training and test datasets. The exhibits are presented by gender to support one of the Challenge's aims of accounting for gender-specific needs. The demographic percentages for years 2 and 3 of the training datasets exclude individuals who had recidivated in the prior year.

Exhibit 1. Challenge Datasets Descriptive Statistics for Females

	Year 1		*Year 2*		*Year 3*	
	Training	Test	Training	Test	Training	Test
Total N	2,217	950	1,760	742	1,396	519
Variable	% *(Recid. %)+*	% *(Recid. %)*	% *(Recid. %)*	% *(Recid. %)*	% *(Recid. %)*	% *(Recid. %)*
Recidivism Overall	20.61 *(NA)*	21.89 *(NA)*	20.68 *(NA)*	17.65 *(NA)*	13.61 *(NA)*	15.06 *(NA)*
Race						
Black	33.51 *(22.21)*	35.68 *(23.01)*	32.84 *(17.47)*	35.18 *(13.79)*	34.17 *(13.63)*	36.82 *(12.89)*
White	66.49 *(19.81)*	64.32 *(21.28)*	67.16 *(22.25)*	64.82 *(19.75)*	65.83 *(13.60)*	63.18 *(16.32)*
Age						
18-22	4.19 *(34.41)*	3.68 *(28.57)*	3.47 *(24.59)*	3.37 *(16.00)*	3.30 *(19.57)*	3.44 *(23.81)*
23-27	15.88 *(25.57)*	17.68 *(26.79)*	14.89 *(25.95)*	16.58 *(23.58)*	13.90 *(14.95)*	15.38 *(22.34)*
28-32	19.40 *(22.33)*	18.53 *(22.73)*	18.98 *(22.75)*	18.33 *(17.65)*	18.48 *(16.28)*	18.33 *(14.29)*
33-37	18.00 *(19.80)*	18.84 *(20.67)*	18.18 *(22.50)*	19.14 *(19.01)*	17.77 *(14.92)*	18.82 *(11.30)*
38-42	14.25 *(19.94)*	13.89 *(20.45)*	14.38 *(20.95)*	14.15 *(16.19)*	14.33 *(11.50)*	14.40 *(18.18)*
43-47	13.13 *(17.53)*	13.16 *(21.60)*	13.64 *(19.17)*	13.21 *(14.29)*	13.90 *(10.82)*	13.75 *(14.29)*
48+	15.16 *(13.69)*	14.21 *(16.30)*	16.48 *(11.72)*	15.23 *(14.16)*	18.34 *(11.33)*	15.88 *(9.28)*

+ Recid. % indicates the percentage of individuals who recidivated by the end of the respective calendar year.

Note: Under each demographic category, the accompanying recidivism percentage indicates the percentage of this category that recidivated by the end of that year. The test dataset was updated throughout the Challenge periods. See the Challenge Background section for more details.

[5] The disclosure risk analysis required collapsing or even completely removing certain demographics that, when included, produced a high risk of re-identification. For example, individuals identified as Asian or Hispanic were removed from the dataset due to the risk of re-identification.

Exhibit 2. Challenge Datasets Descriptive Statistics for Males

	Year 1		*Year 2*		*Year 3*	
	Training	Test	Training	Test	Training	Test
Total N	15,811	6,857	10,891	4,781	8,002	3,535
Variable	% *(Recid. %)+*	% *(Recid. %)*	% *(Recid. %)*	% *(Recid. %)*	% *(Recid. %)*	% *(Recid. %)*
Recidivism Overall	31.12 *(NA)*	31.19 *(NA)*	26.53 *(NA)*	25.07 *(NA)*	20.01 *(NA)*	20.65 *(NA)*
Race						
Black	60.53 *(31.69)*	61.18 *(32.16)*	60.02 *(26.45)*	60.32 *(25.02)*	60.08 *(20.45)*	60.37 *(20.29)*
White	39.47 *(30.24)*	38.82 *(29.68)*	39.98 *(26.64)*	39.68 *(25.07)*	39.92 *(19.35)*	39.63 *(21.20)*
Age						
18-22	8.61 *(42.47)*	8.41 *(47.83)*	7.19 *(35.39)*	6.38 *(30.23)*	6.27 *(26.10)*	5.94 *(26.19)*
23-27	20.61 *(37.28)*	20.37 *(36.65)*	18.77 *(32.88)*	18.76 *(29.04)*	17.15 *(24.27)*	17.77 *(26.59)*
28-32	19.09 *(33.82)*	19.79 *(31.91)*	18.35 *(29.03)*	19.58 *(28.57)*	17.72 *(22.71)*	18.67 *(22.42)*
33-37	16.29 *(29.97)*	16.29 *(28.02)*	16.56 *(25.83)*	17.04 *(25.75)*	16.72 *(21.23)*	16.89 *(23.12)*
38-42	10.90 *(27.32)*	11.97 *(28.14)*	11.50 *(24.02)*	12.51 *(19.32)*	11.90 *(19.54)*	13.47 *(17.44)*
43-47	9.91 *(26.16)*	9.29 *(24.65)*	10.62 *(22.64)*	10.17 *(23.13)*	11.18 *(16.20)*	10.44 *(18.97)*
48+	14.58 *(19.65)*	13.87 *(22.82)*	17.00 *(17.66)*	15.56 *(18.94)*	19.06 *(13.11)*	16.83 *(11.60)*

+ Recid. % indicates the percentage of individuals who recidivated by the end of the respective calendar year.

Note: Under each demographic category, the accompanying recidivism percentage for this category is presented. The test dataset was updated throughout the Challenge periods. See the Challenge Background section for more details.

Variables Provided for Each Challenge Period

The training and test datasets were both released at the initial launch of the Challenge. The training dataset contained the previously stated independent variables for the three years of the data and four dichotomous dependent variables measuring recidivism across the three-year supervision period. These dependent variables indicated whether an individual recidivated at any time in the three-year follow-up period (yes/no) and whether they recidivated in year 1, year 2, or year 3 (yes/no for each). The test dataset only included the independent variables (with the exception of supervision activities, which were included starting in year 2) and was updated twice during the Challenge. The Challenge consisted of three submission periods. The first Challenge period opened on April 30th, 2021, and closed on May 31st, 2021. The second Challenge period opened on June 1st, 2021, and closed on June 15th, 2021. The third Challenge period opened on June 16th, 2021, and closed on June 30th, 2021. See exhibit 3 for the variables included in each Challenge period's dataset.

Exhibit 3. Variables Provided in Challenge Datasets Throughout the Challenge Period

	Challenge Datasets			
Challenge Variable Domains	Training Dataset	Test Dataset Year 1	Test Dataset Year 2	Test Dataset Year 3
Supervision Case Information	X	X	X	X
Prison Case Information	*X*	*X*	*X*	*X*
Prior Georgia Criminal History	X	X	X	X
Prior Georgia Community Supervision History	X	X	X	X
Conditions of Current Supervision Period	X	X	X	X
Supervision Activities for Current Supervision Period	X		X	X
Recidivism Variables	X			

Note: The test dataset was updated throughout the Challenge periods. See the Challenge Background section for more details.

In the first Challenge period, submissions forecasted the percentage likelihood that each individual within the test dataset would recidivate within the first year of release. At the conclusion of the first period, the following changes were made to the datasets:

- The test dataset was updated by removing individuals who recidivated during year 1.
- Supervision variables were added to the test dataset.

The supervision activity variables described which activities individuals engaged in over the course of their remaining supervision (e.g., average number of delinquency reports per year, average number of program attendances per year). As shown in exhibit 3, the supervision activity variables were not provided for the first period, mirroring real-world circumstances in which supervision officers do not initially know this information when attempting to identify whether an individual will recidivate.

In the second Challenge period, submissions forecasted the likelihood that individuals remaining in the test dataset would recidivate by the end of year 2. At the conclusion of the second period, the test dataset was then updated again by removing those who recidivated. However, the supervision activity variables were not updated because the values for supervision were an average of individuals’ engagement in each item over the three years of supervision. In the third Challenge period, submissions then forecasted whether the individuals who remained in the dataset would recidivate by the end of year 3.

Challenge Prize Structure

The Challenge contained three categories (student, small business, and large business), each with its own prize structure.[6] Eligibility for the student category was limited to full-time high school and undergraduate students. The small business category comprised teams of one to 10 individuals (e.g., graduate students, professors, or other individuals) and small businesses with fewer than 11 employees. The large business category was designated for businesses with 11 or more employees. Entrants could choose to submit to a higher participant category. The results of the small and large business categories are included in this paper.[7] Across the three Challenge periods, NIJ received and scored 57 entries for the year 1 forecast, 55 for the year 2 forecast, and 54 for the year 3 forecast. Teams were allowed to provide predictions for more than one Challenge period.

[6] "Recidivism Forecasting Challenge: IX. Prizes: Student Category," *National Institute of Justice*, https://nij.ojp.gov/funding/recidivism-forecasting-challenge#student-category.

[7] Student winners were not included in this analysis because the submissions performed substantially worse than small and large team winners. Allowing teams of students to submit into this category may have improved the results.

Judging Criteria

In order to meet the Challenge's aims of improving recidivism forecasts by identifying key variables, addressing gender differences, and accounting for racial bias, the categories and scores below were examined to select Challenge winners. For each Challenge period, submissions were judged on (1) the accuracy of their recidivism forecasts for males and females, and the average of these two accuracy scores, and (2) the fairness of their recidivism forecast accuracy when accounting for racial bias between Black and white individuals on parole, for both males and females.

To measure model accuracy, NIJ calculated the mean squared error of entries using the following Brier score:[8]

$$Brier\ score = \frac{1}{n}\sum_{i=1}^{n}(f_i - A_i)^2$$

In this calculation, n is the count of individuals in the test dataset and f_i is the forecasted probability of recidivism [0, 1] for individual i. A_i is the actual outcome {0, 1} for individual i, where 0 indicates individual i did not recidivate and 1 indicates they did recidivate. Three separate accuracy prizes were awarded: Brier score among females, Brier score among males, and a combined-gender Brier score that was an equally weighted composite of the male and female Brier scores. As the Brier score is a measure of error, submissions with the lowest scores won prizes.

To measure accuracy while accounting for fairness, NIJ modified each submission's Brier score, as calculated above, with a penalty based on the difference in false positive rates (FPR) (i.e., the number of individuals who are forecasted to recidivate, but do not recidivate) between white and Black individuals. To do this, NIJ used a 0.5 threshold to

[8] Glenn W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review* 78 no. 1 (1950): 1-3, https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2.

convert the entries' predicted probabilities into binary (i.e., yes/no) predictions. The fairness penalty (FP) function, crafted by NIJ for the Challenge, was the difference in false positive rates between individuals of different racial groups in a given year's test dataset:

$$FP = 1 - |FPR_{Black} - FPR_{white}|$$

Because the Brier score measures the error of predictions and is bound by 0 and 1, NIJ considered 1 minus the Brier score a metric of accuracy. The index NIJ used to calculate which algorithms were the most accurate while accounting for racial bias was:

$$Fair\ and\ Accurate = (1 - Brier\ score)(FP)$$

As this is now a measure of fairness and accuracy, submissions looked to maximize this metric (i.e., the highest scores won). Although a false negative rate or a combination of false negative and false positive rates could have been selected, NIJ selected a false positive rate as the metric on which to evaluate submissions as this would result in an individual being labeled as high risk when he or she is not. The consequences of incorrectly being identified as high risk could lead to excessive supervision services (i.e., additional supervision requirements and potentially additional service requirements), which have been linked to negative outcomes for those under supervision.[9] Assigning undue supervision requirements also means fewer supervision resources for those who may benefit from supervision.

NIJ awarded separate accuracy prizes for each category (i.e., student, small business, and large business). The prizes for fairness and accuracy were awarded in all categories for both the top scores forecasting males and the top scores forecasting females. Therefore, all submissions competed with one another for fairness and accuracy prizes. The winning submissions and their individual scores can be found on NIJ's Recidivism Forecasting Challenge Official Results webpage.[10]

Models and Methods Used for Contextualizing and Comparison

To put into context the winning Brier scores, it is necessary to know the accuracy of a random chance model and other simple models that slightly improve forecasting accuracy. The simplest model for determining who is likely to recidivate within the next year is to assign everyone a 50% probability. This likelihood would essentially be flipping a coin for every person (i.e., heads for recidivating in the next year, tails for not recidivating in the next year). If we consider recidivating equal to 1 and not recidivating equal to 0, and we assign everyone a 50% probability (0.5), the Brier score for this random chance model would equal 0.25 because the error for each individual is 0.5 (i.e., |1 – 0.5| = 0.5 or |0 – 0.5| = 0.5 for all individuals); averaging and squaring the error obtains a final Brier score of 0.25 (i.e., 0.52 = 0.25). The more accurate a model is, the lower its associated Brier score is. Therefore, at a minimum,

[9] Christopher T. Lowenkamp, Edward J. Latessa, and Alexander M. Holsinger, "The Risk Principle in Action: What Have We Learned From 13,676 Offenders and 97 Correctional Programs?" *Crime & Delinquency* 52 no. 1 (2006): 77-93, https://doi.org/10.1177%2F0011128705281747.

[10] "Recidivism Forecasting Challenge: Official Results," *National Institute of Justice,* July 28, 2021, https://nij.ojp.gov/funding/recidivism-forecasting-challenge-results.

NIJ would want any model used in practice to have a better forecasting accuracy than flipping a coin — i.e., a Brier score less than 0.25.

In addition to the random chance model, NIJ created several naive models based on the descriptive statistics of the corresponding years' training dataset. Naive models were calculated based on recidivism for the overall population and used simple demographics (race, gender, and age) as single-factor, two-factor, and three-factor models. These simple, naive models provide another standard (other than random chance) for comparing how well the winning forecasts performed. Eight naive models for each of the three Challenge periods were calculated, based on the average recidivism rates of:

1. The overall population
2. White and Black individuals
3. Females and males
4. Each five-year age group between ages 18 and 48+
5. A mutually exclusive combination of models 2 and 3
6. A mutually exclusive combination of models 2 and 4
7. A mutually exclusive combination of models 3 and 4
8. A mutually exclusive combination of models 2, 3, and 4

The average recidivism rates for the naive demographic models were calculated from the training dataset. For each of the eight models, an individual received a single probability of recidivating in the next year based on his or her demographic information. For example, the third naive model assigned all females a recidivism probability of 0.2061 and all males a probability of 0.3112 in year 1 because these were the average recidivism rates in the year 1 training dataset for females and males, respectively. The specific probabilities for naive models 3, 5, and 7 can be found in exhibits 1 and 2. These and the remaining naive model probabilities for each year can also be found in the appendix.

The Brier score is used in the literature to determine the accuracy of a given model; however, it is not a good measure for comparison of model accuracy. To contextualize relative model accuracy — specifically, how the accuracies of the winning models compared to the random chance model — NIJ used the Brier skill score.[11] The Brier skill score for a model is calculated by taking 1 minus the Brier score of that model and dividing by the Brier score of a base model, as follows:

$$Brier\ skill\ score = 1 - \frac{Brier\ score}{Brier\ score\ of\ base\ model}$$

[11] C.A.T. Ferro and T.E. Fricker, "A Bias-Corrected Decomposition of the Brier Score," *Quarterly Journal of the Royal Meteorological Society* 138 no. 668 (2012): 1954-1960, https://doi.org/10.1002/qj.1924.

The Brier skill score is a number in the range $(-\infty, 1]$. A model with lower accuracy than the base model will return a Brier skill score less than 0, a model with the same accuracy as the base model will return a Brier skill score of 0, and a model with perfect accuracy will return a Brier skill score of 1. The Brier skill score is, therefore, a relative metric that examines how two models perform compared to one another. The following sections will explore the improvements in accuracy obtained from using naive demographic models over a random chance model, how well the winning models compared to the best- and worst-performing naive models, as well as their performance on the fairness and accuracy measure defined above.

Results: Naive Models

This section compares the naive models' Brier scores among males and females in the dataset for years 1 to 3. Exhibit 4 summarizes the Brier scores for each of the naive models. The poorest performing naive demographic model in each column is noted with a dagger, while the best performing model is noted with a double dagger.

Exhibit 4. Male and Female Brier Scores of Naive Models for Years 1 to 3

(Model) Demographic Categories	Female			Male		
	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
Random Chance	0.25000	0.25000	0.25000	0.25000	0.25000	0.25000
(1) Overall Population	0.19606†	0.15187*	0.12950	0.21482†	0.18791*	0.16411
(2) Race	0.19550	0.15187*†	0.12959†	0.21465	0.18791*	0.16425†
(3) Gender	0.19290	0.14630	0.12811*‡	0.21464	0.18808*†	0.16390
(4) Age	0.19316	0.15126	0.12862	0.21061	0.18675	0.16189*
(5) Race/Age	0.19291	0.15120	0.12879	0.21008	0.18647‡	0.16189*
(6) Gender/Age	0.19049	0.14617	0.12854	0.21059	0.18691	0.16168
(7) Gender/Race	0.19281	0.14545	0.12811*	0.21451	0.18808*	0.16398
(8) Gender/ Race/Age	0.19015‡	0.14506‡	0.12860	0.21001‡	0.18664	0.16163‡

* Indicates a difference beyond five significant digits.
† Indicates the worst Brier scores.
‡ Indicates the best Brier scores.

Exhibit 4 shows that, based on descriptive statistics, all of the naive models were more accurate and had lower Brier scores than the random chance model, for every year and for both males and females. The information in exhibit 4 can be used to calculate the Brier skill score between two models. For example, when comparing the "best" to the "worst" naive demographic models, the resulting Brier skill scores ranged from 0.009 to 0.045 for a given prize category. This means that the percentage improvements among the naive demographic models were small (i.e., less than 5%).

Exhibit 5 shows the Brier skill score of each of the naive demographic models compared to the random chance model. This table was calculated using the Brier skill score equation and the information in exhibit 4, using the random chance model as the base model for each calculation. The average Brier skill scores of models among females are 0.228, 0.405, and 0.485 in each consecutive year. Among males, the average Brier skill scores are, respectively, 0.150, 0.251, and 0.348. Therefore, the models' accuracy relative to the random chance model increases between years 1, 2, and 3 across both genders. This makes sense, as the random chance model consistently predicts a 0.5 probability of recidivating, while the overall recidivism rates in the population of our training datasets decrease each year.

Exhibit 5. Male and Female Brier Skill Scores of Naive Models Compared to Random Chance Model for Years 1 to 3

(Model) Demographic Categories	Female			Male		
	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
(1) Overall Population	0.21576	0.39252	0.48200	0.14072	0.24836	0.34356
(2) Race	0.21800	0.39252	0.48164	0.14140	0.24836	0.34300
(3) Gender	0.22840	0.41480	0.48756	0.14144	0.24768	0.34440
(4) Age	0.22736	0.39496	0.48552	0.15756	0.25300	0.35244
(5) Race/Age	0.22836	0.39520	0.48484	0.15968	0.25412	0.35244
(6) Gender/Age	0.23804	0.41532	0.48584	0.15764	0.25236	0.35328
(7) Gender/Race	0.22876	0.41820	0.48756	0.14196	0.24768	0.34408
(8) Gender/Race/Age	0.23940	0.41976	0.48560	0.15996	0.25344	0.35348
Average Brier Skill Score	0.22801	0.40541	0.48507	0.15005	0.25063	0.34834

Note: This table shows the Brier skill scores comparing each of our eight naive demographic models to the random chance model. Brier skill score values were calculated from the Brier scores in exhibit 4.

The Brier scores generally decreased as the naive models became more complex (i.e., incorporated more demographic information), but there were exceptions. For males, of the three demographic criteria considered (gender, race, and age), categorizing by age was the most important factor as indicated by the largest Brier score improvement (model 4), versus gender (model 3) or race (model 2). The age-only model (4) was further improved for males when also considering race (model 5) and showed minimal additional improvement when also accounting for gender (model 8). Therefore, for males in our dataset, the most accurate naive models appeared to be the race/age model (5) and the gender/race/age model (8).

For females, of the three demographic criteria considered, categorizing by gender was the most important factor with the largest Brier score improvement (model 3), versus adding information on age (model 2) or race (model 4). The gender-only naive model (3) was the most accurate model for year 3. In years 1 and 2, the accuracy was further improved when also considering age (model 6) and showed minimal additional improvement when also accounting for race (model 8). Therefore, the most accurate naive models among females for this sample appeared to be the gender, gender/age, and gender/race/age models.

Results: Brier Score Winners

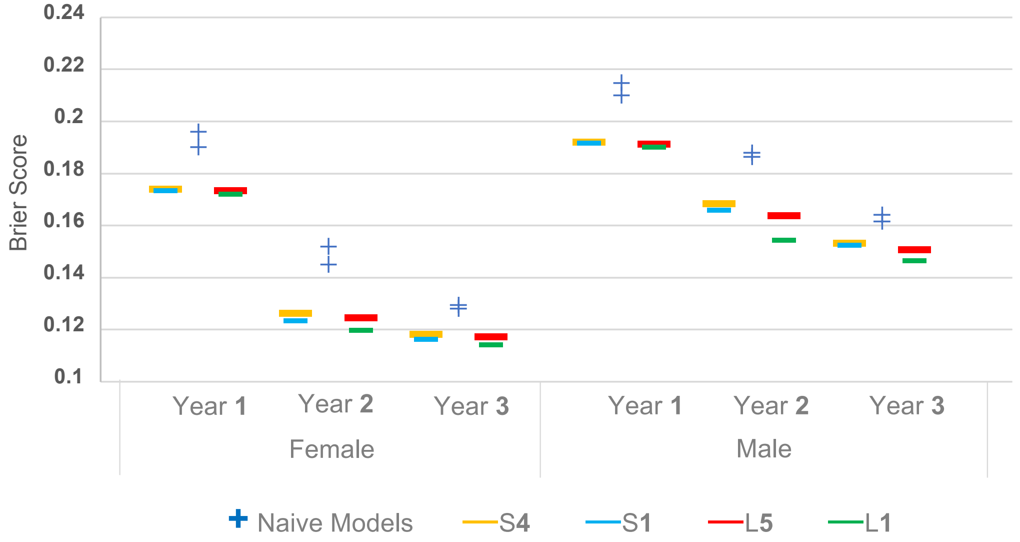
This section compares the winning models to the best naive models in terms of accuracy (i.e., Brier score, and Brier skill score compared to the random chance model). The "best" naive model refers to the model with the lowest Brier score for the respective year and gender group, as noted in exhibit 4, while the "worst" naive model refers to the highest-scoring naive demographic model, also noted in exhibit 4. Exhibit 6 shows the winning Brier scores, ranging from fourth-place small businesses (S4) to first-place large businesses (L1), for each year among males and females, along with the random chance model and the best and worst naive demographic models. Exhibit 7 plots the values of exhibit 6 to display the range and trend of accuracy between the naive and winning models across all years and genders.

Exhibit 6. Male and Female Brier Scores of Winning Models vs. Naive Models for Years 1 to 3

	Female			Male		
	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
Random Chance	0.25000	0.25000	0.25000	0.25000	0.25000	0.25000
Worst Naive	0.19606	0.15187	0.12959	0.21482	0.18808	0.16425
Best Naive	0.19015	0.14506	0.12811	0.21001	0.18647	0.16163
S4	0.17400	0.12630	0.11820	0.19220	0.16850	0.15340
S1	0.17330	0.12330	0.11614	0.19160	0.16580	0.15237
L5	0.17340	0.12450	0.11734	0.19140	0.16380	0.15070
L1	0.17190	0.11960	0.11390	0.19000	0.15420	0.14630

Note: The "best" and "worst" naive models are taken from exhibit 4 and vary depending on which naive demographic models had the lowest and highest Brier scores, respectively, for a given year and gender. Five large businesses and four small businesses were awarded prizes for each series. The exhibit shows Brier scores for the small businesses in fourth (S4) and first (S1) places and the large businesses in fifth (L5) and first (L1) places.

Exhibit 7. Naive vs. Winning Models: Highest and Lowest Brier Scores for Each Accuracy Prize Category



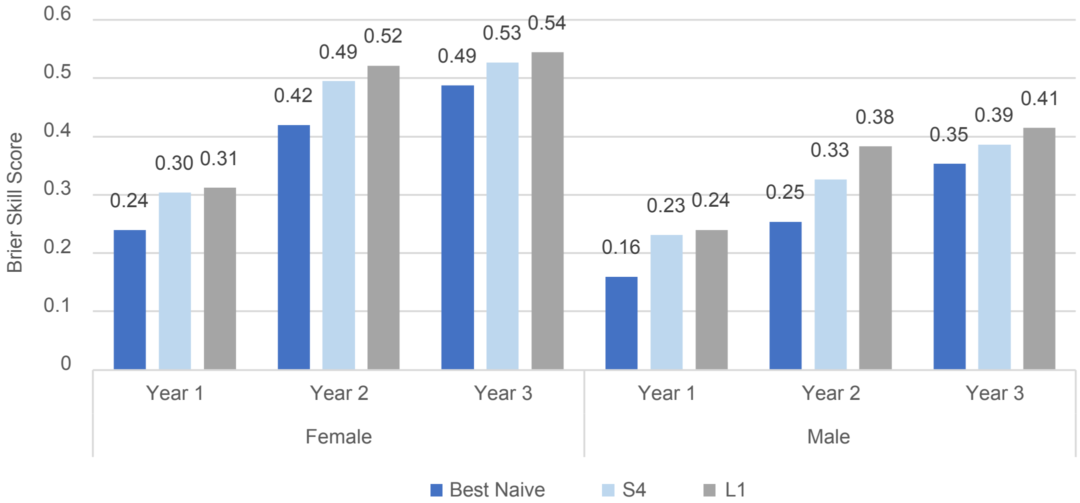
Note: Brier scores from select naive and winning forecasts are presented to display the range of winning scores across years and show how these scores compare to the best and worst naive models. The naive models with the highest and lowest Brier scores (see exhibit 6) for males and females across years 1 to 3 are presented along with select small and large business winners. Because five large businesses and four small businesses were awarded prizes for each series, the exhibit shows Brier scores for the large businesses in first (L1) and fifth (L5) places and the small businesses in first (S1) and fourth (S4) places.

Exhibit 6 shows that the range of Brier scores was relatively small among winners in each accuracy prize category. These ranges from exhibit 6 are also graphed in exhibit 7. Additionally, NIJ calculated the Brier skill score between the L1 winner and the S4 winner, where the S4 winner was used as the base model for each prize category (e.g., Female — Year 3). To calculate the Brier skill score, NIJ used the L1 and S4 Brier score values from exhibit 6 to compare the relative Brier score improvement. For a given prize category, the corresponding Brier skill score ranged from 0.011 to 0.085 when comparing the L1 to the S4 submissions. Male — Year 2, with a Brier skill score of 0.085, was the only prize category to have a Brier skill score greater than 0.05 when comparing the L1 to the S4 Brier scores. This category also has the largest visible difference between L1 and S4, as seen in exhibit 7.

As shown in exhibit 7, the accuracy of models improved as the years progressed via the decrease in Brier scores for each year. This trend is consistent across naive models and winning models for both females and males. Additionally, the improvement in Brier scores between years 2 and 3 appears smaller than the improvement between years 1 and 2.

Exhibit 8 shows the Brier skill score of the best naive model and S4 and L1 winning models, compared to the random chance model. Across all three years, all models scored better than

Exhibit 8. Brier Skill Scores Compared to Random Chance Model



Note: This exhibit graphs the Brier skill scores of the best naive model, the fourth-place small business model (S4), and the first-place large business model (L1) when using the random chance model as the base model for comparison of all three.

the random chance model (i.e., Brier skill scores greater than 0). Again, Brier skill scores were larger for females and for later years. Generally, the best naive demographic models had the lowest Brier skill score over chance, while the L1 winner had the highest Brier skill score. Therefore, the winning models — especially the large business winners — were more accurate than the naive models based on simple static demographics. Additionally, exhibits 7 and 8 show that the Brier scores for females were lower than the scores for males in a given year, and Brier skill scores for females were higher than for males. This trend exists across the naive demographic models and winning models and is most prominent in years 2 and 3.

Results: Fairness and Accuracy Prize Winners

Scores for fairness and accuracy were calculated for each submission as described in the Challenge Background section. NIJ did not compare the naive models to the winning models here because none of the naive models received a fairness penalty to their Brier scores. This was because all naive models produced recidivism forecasts less than 0.5, meaning that everyone was predicted not to recidivate (i.e., resulting in no false positives), as described in the Judging Criteria section. Since NIJ only penalized false positives, none of the naive models received a fairness penalty.

There were five winners for each of the Challenge's six fairness and accuracy prize categories. Exhibit 9 summarizes the number of submissions and winners who received fairness penalties and the average penalty assessed. Almost every prize category had multiple winning submissions that received a penalty, except for the Female — Year 3 prize category. If a winning model received a penalty, it was small; all winners had penalties lower than 0.005, which was less than half of the average penalty assessed across all submissions.

Exhibit 9. Summary of Fairness Penalties

	Female			Male		
	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
Submissions With Penalties	47/57	44/55	28/54	50/57	49/55	41/54
Winners With Penalties	2/5	2/5	0/5	4/5	3/5	4/5
Average Submission Penalty	0.0163	0.0196	0.0251	0.0183	0.0159	0.0122
Average Winner Penalty	0.0019	0.0016	0.0000	0.0008	0.0034	0.0018

Note: Winners are submissions that were awarded prizes for fairness and accuracy.

Exhibit 10. Percentages of Submissions Penalized for Fairness

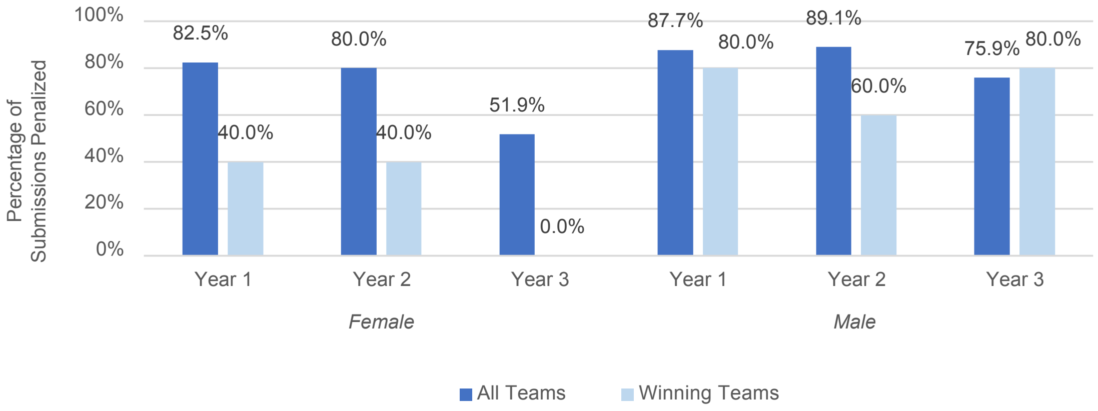
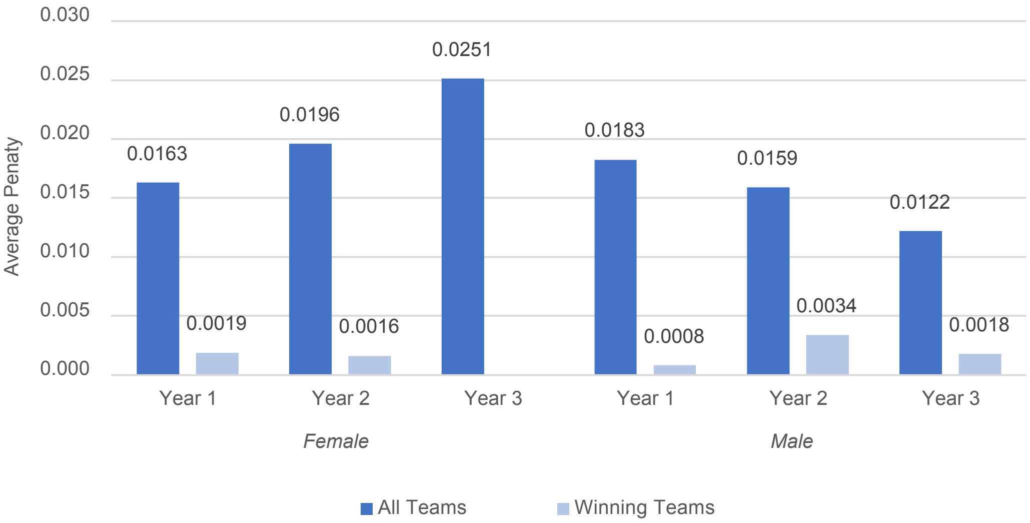


Exhibit 11. Average Fairness Penalties for Penalized Submissions



Note: Averages are calculated only to include submissions that receive a penalty. See exhibit 9 for the numbers of winning submissions and submissions overall that were and were not penalized.

Exhibit 10 shows the percentages of all submissions and winning submissions that were penalized in each prize category. A majority of all submissions in each prize category received a fairness penalty. The number of submissions penalized decreased from years 1 to 3 when predicting female recidivism, but not when predicting male recidivism. Additionally, the number of submissions with a fairness penalty was higher for males than for females across all years. Winning submissions received fewer penalties for predicting female recidivism compared to all submissions A similar pattern was not seen for predicting male recidivism; winning submissions were penalized at around the same rate as submissions overall. However, as shown in exhibits 9 and 11, the average magnitude of those penalties was much lower for winning submissions than for submissions overall. This suggests that winning submissions were able to predict recidivism among females more fairly. This is likely because recidivism rates in our dataset were lower for females and the same recidivism prediction threshold of 0.5 was used for both genders.

The average penalty of submissions that received a fairness penalty across gender and years is presented in exhibit 11. The average overall fairness penalty increased for females as the years progressed, but the opposite was true for males. Meanwhile, the winning submissions had lower or no penalties when predicting recidivism for females across the years. There was no clear trend for winners' penalties when predicting recidivism among males. The average winners' penalty was between 78.62% and 100%, which was lower than the average penalty across all submissions, as shown in exhibit 11. This means that certain methods or prediction variables may be more prone to racial bias (i.e., a difference in false positive rates across races).

Conclusion

This paper presents the initial results from NIJ's Fiscal Year 2021 Recidivism Forecasting Challenge and provides context as to the level of accuracy and fairness reflected in the Challenge winners' scores as they forecast who would recidivate. With the Challenge concluded, NIJ is seeking to encourage discussion on reentry, bias, fairness, measurement, and algorithm advancement.

Except for the fairness and accuracy scores, the top winners in each prize category were compared against a random chance model and eight additional naive models. These naive models calculated the average recidivism in each year based on the overall population of the dataset and combinations of demographic information (i.e., race, gender, and age). NIJ used Brier scores to measure a model's error rate and the Brier skill score to compare two models' accuracy. In each prize category for both small and large businesses, the winning scores had lower error rates than the naive models. The differences in accuracy between the winning and naive models may be attributed to winning submissions incorporating additional information (e.g., census data) and/or using more advanced techniques (e.g., regression, random forest, neural networks). The methods and information used to craft the winning models are currently being reviewed, and papers from the winning submissions describing their approach will be made available at a future date. Additionally, NIJ plans to release papers exploring and summarizing the factors responsible for the increased accuracy observed in the winning submissions.

Scores were also compared to assess fairness and accuracy. If, for example, a model did not have the same false positive rates for white and Black males, it received a penalty to its male fairness and accuracy score; the same was done for females. The results show there were penalties across the winning submissions, although they were considerably smaller than the average penalty assessed across all submissions. Further exploration is needed to understand better how this fairness was reflected in the Challenge and what implications these fairness results have for the field.

The results from the Challenge suggest several areas for further study. Future studies might consider what data can be added to common risk assessment tools that are currently in use (e.g., LSI, Compass, ORAS) so that they can provide more accurate recidivism forecasts. To test this, agencies could use available information to better understand who is at higher risk of recidivating. Further examination is also needed to identify and understand gender differences in risk assessments and the support provided for individuals under community supervision. Additionally, it is worth noting that the majority of the winners' forecasts received fairness penalties, albeit ones smaller than the average penalties for all submissions. Further work is needed to unpack these penalty scores along with continued discussions on the proper balance between fair and accurate forecasts. NIJ intends to address these research questions and examine other related issues in the future.

Acknowledgments

NIJ would like to acknowledge our partners at the Georgia Department of Community Supervision for providing the dataset, NIJ staff who assisted with the Challenge who are not authors of this paper, our colleagues at the Bureau of Justice Statistics and Bureau of Justice Assistance for their assistance in developing the Challenge, and the Challenge entrants for their participation.

Appendix: Average Recidivism Rates for Naive Models

Eight naive models for each year were constructed using the training dataset. The first naive model 1 looked at the average recidivism rate across all individuals. The models 2, 3, and 4 looked at the average recidivism rates for Blacks compared to whites, males compared to females, and each of the age groups, respectively. Models 5, 6, and 7 considered the average number of individuals who recidivate with respect to two demographics (e.g., Black males, white males, Black females, and white females). The final naive model 8 was constructed by looking at the average number of individuals who recidivate when you have information about all three demographics (e.g., Black male age 18-22, Black male age 23-27, …, Black male age 48+). Exhibit A shows the averages for each of the three years for all eight models that were generated.

Exhibit A. Average Fairness Penalties for Penalized Submissions

Naive Model	Naive Model Demographic Composition		Average Recidivism Rates		
			Year 1	Year 2	Year 3
Model 1 (Overall Population)	No Information		.2983	.2571	.1906
Model 2 (Race Only)	Black		.3101	.2572	.1983
	White		.2824	.2570	.1806
Model 3 (Gender Only)	Male		.3112	.2653	.2001
	Female		.2061	.2068	.1361
Model 4 (Age Only)	18-22		.4195	.3507	.2555
	23-27		.3614	.3209	.2312
	28-32		.3239	.2813	.2172
	33-37		.2861	.2533	.2024
	38-42		.2618	.2351	.1814
	43-47		.2481	.2205	.1524
	48+		.1889	.1685	.1286
Model 5 (Race and Age)	Black	18-22	.4286	.3639	.2701
		23-27	.3646	.3181	.2323
		28-32	.3203	.2755	.2138
		33-37	.2775	.2318	.2092
		38-42	.2633	.2021	.1695
		43-47	.2715	.2182	.1628
		48+	.2138	.1832	.1413
	White	18-22	.3976	.3203	.2241
		23-27	.3555	.3260	.2291
		28-32	.3286	.2892	.2219
		33-37	.2967	.2809	.1931
		38-42	.2602	.2675	.1942
		43-47	.2258	.2225	.1431
		48+	.1613	.1533	.1159
Model 6 (Gender and Age)	Male	18-22	.4247	.3589	.2610
		23-27	.3728	.3288	.2427
		28-32	.3382	.2903	.2271
		33-37	.2997	.2583	.2123
		38-42	.2732	.2402	.1954
		43-47	.2616	.2264	.1620
		48+	.1965	.1766	.1311
	Female	18-22	.3441	.2459	.1957
		23-27	.2557	.2595	.1495
		28-32	.2233	.2275	.1628
		33-37	.1980	.2250	.1792
		38-42	.1994	.2095	.1150
		43-47	.1753	.1917	.1082
		48+	.1369	.1172	.1133

Exhibit A. Average Fairness Penalties for Penalized Submissions (continued)

	Naive Model Demographic Composition		Average Recidivism Rates		
			Year 1	**Year 2**	**Year 3**
Model 7 (Gender and Race)	Black Male		.3169	.2645	.2045
	White Male		.3024	.2664	.1935
	Black Female		.2221	.1747	.1363
	White Female		.1981	.2225	.1360
Model 8 (Gender, Race, and Age)	Black Male	18-22	.4315	.3732	.2792
		23-27	.3717	.3295	.2397
		28-32	.3281	.2781	.2228
		33-37	.2829	.2393	.2161
		38-42	.2646	.2065	.1766
		43-47	.2817	.2260	.1681
		48+	.2196	.1866	.1387
	White Male	18-22	.4069	.3229	.285
		23-27	.3751	.3273	.2489
		28-32	.3540	.3101	.2343
		33-37	.3241	.2877	.2059
		38-42	.2830	.2800	.2198
		43-47	.2401	.2269	.1558
		48+	.1675	.1647	.1225
	Black Female	18-22	.3636	.1786	.1304
		23-27	.2585	.1743	.1556
		28-32	.2148	.2453	.1125
		33-37	.2126	.1500	.1412
		38-42	.2500	.1594	.1034
		43-47	.1828	.1579	.1250
		48+	.1429	.1444	.1688
	White Female	18-22	.3265	.3030	.2609
		23-27	.2537	.3203	.1442
		28-32	.2271	.2193	.1854
		33-37	.1912	.2591	.1534
		38-42	.1786	.2283	.1197
		43-47	.1717	.2073	.1000
		48+	.1342	.1050	.0894