These were our initial variables in the one-year forecasting model. Shortly before the deadline, we attempted an enhancement. We used these variables to also predict two-year recidivism and three-year recidivism. We developed out-of-sample forecasts for each of these variables. Then we predicted these variables for the test set as well. This is a process called stacking. We then used these forecasts for two-year and three-year recidivism along with the variables discussed previously to predict one-year recidivism. Our attempt with this multi-stage model was to add information about two-year and three-year recidivisms to increase the sample size of data available to fit the model. Since this model performed slightly better for a short validation exercise before the deadline, we used the forecast with stacking. But we chose to use the concept behind the simpler model for the second and third rounds of the competition after further validation analysis. Table 1 shows the performance for the one-year forecasts on the test set for the model used in the submission and the simpler model.

Table 1. Brier score of simple model and stacked model for one-year forecast

	Male	Female
Simple Model	0.191185	0.155625
Stacked Model (Submission)	0.191306	0.155214

A measure of importance for variables in a gradient boosting model is the relative contribution of each variable in the model. These values are calculated by measuring the decrease in model performance by removing variables and normalizing the sum to 100%. Table 2 shows the relative importance of variables in the submission for the one-year forecast. We see age and gang affiliation are the two most influential variables in the model. The prediction for two-year recidivism is the next most significant variable, but this variable is derived as a prediction from other variables in the model.

Table 2. Relative importance of variables for the one-year forecast for the submission.

Variable	Relative Importance
Age_at_Release	14.9%
Gang_Affiliated	9.7%
Probability Recidivism Two Years	8.1%
Total Significant Variables	6.6%
Prison_Years	5.0%
Supervision_Risk_Score_First	4.7%
Prior_Arrest_Episodes_Property	4.5%
Prison_Offense	4.1%
Residence_PUMA	4.0%
Total Arrests	3.7%
Prior_Arrest_Episodes_PPViolationCharges	3.6%
Education_Level	3.5%
Prior_Arrest_Episodes_Felony	3.1%
Supervision_Level_First	3.1%
Condition_MH_SA	2.5%
Probability Recidivism Three Years	2.1%
Prior_Conviction_Episodes_Misd	1.7%
Prior_Revocations_Parole	1.5%
Prior_Conviction_Episodes_Drug	1.5%
Prior_Conviction_Episodes_Prop	1.4%
Total Convictions	1.4%
Prior_Arrest_Episodes_Misd	1.4%
Race	1.3%
Dependents	1.0%
Condition_Cog_Ed	0.9%
Prior_Arrest_Episodes_Violent	0.9%
Gender	0.9%
Prior_Conviction_Episodes_Felony	0.6%
Prior_Conviction_Episodes_PPViolationCharges	0.5%
Prior_Arrest_Episodes_GunCharges	0.4%
Prior_Revocations_Probation	0.3%
Condition_Other	0.3%
Prior_Conviction_Episodes_Viol	0.3%
Prior_Arrest_Episodes_DVCharges	0.3%
Prior_Conviction_Episodes_DomesticViolenceCharges	0.2%
Prior_Conviction_Episodes_GunCharges	0.1%

Table 3 shows the confusion matrix for the one-year forecast for the submission. This confusion matrix is based on a threshold of 0.5. Since the model is optimizing squared error and the portion of parolees who are recidivists is less than 0.5, the number of predicted recidivists is low relative to those predicted not to be recidivists.