PREDICTING RECIDIVISM

Final Report
December 2021

Thomas L Champion

Table of Contents

Abstract

Compared to it's worldwide peers, the United States has an unusually high incarceration rate. Reducing the level of recidivism is one factor that can lower this rate. There are, however, many challenges in doing so, one of which is being able to most effectively use the limited resources available. This can be aided by determining which individuals are most likely to experience recidivism and prioritizing resources based upon this information. This project aimed to develop a machine learning framework to identify these individuals based upon a range of person- and location-based variables. Results found that XGBoost-based model provided the best predictive capability, with a Brier Score of 0.1922 and a ROC-AUC score of 0.7653.

Introduction

Currently, the United States has the highest incarceration rate worldwide, with 655 individuals incarcerated per 100,000 people[8]. This far surpasses the worldwide rate of 145 per 100,000[8]. This high level of incarceration has both a social and economic impact on society. In 2010 it was estimated that cutting the prison population in half would save $16.9 billion dollars annually, with most of those savings being seen by state or local governments[6]. This savings could be redirected into other social services, which may in itself help to lower crime levels and decrease the pool of potential inmates[1].

One method of reducing the prison population is by ensuring that those individuals released from prison do not experience recidivism. Currently in the United States, 76.6% of released prisoners will be rearrested within 5 years[2]. While there are many options for reducing this rate one limitation in any such effort is the availability of resources, both personnel and financial. The aim of this project is to develop a machine learning framework which can predict an individual's likelihood of experiencing recidivism. In so doing, limited resources can be targeted most effectively to provide those services and support aimed at reducing an individual's likelihood of being returned to the prison system.

Related Work

The data used for this project was also used by the National Institute for Justice in their Recidivism Forecasting Challenge[5], which aimed to improve the ability to forecast recidivism based upon a number of person- and place-based variables. Data was made available based upon recidivism within 1,2 and 3 years of release from prison and entries were divided into 3 groups: student, small team and large team. A total of 172 submissions were made, with minimum Brier Scores of 0.13 for the 3 year recidivism category.

Data

Data Description

The data set used was obtained from the National Institute of Justice, consisting of data from the State of Georgia about individuals released from prison to parole from January 1, 2013 to December 31, 2015. There are a total of 25,835 observations with 54 columns in the data set. The full data dictionary can be found in Appendix A.

The value of Recidivism_Within_3years was chosen as the target variable, indicating if the individual experienced recidivism within 3 years of release from prison. 5 columns were removed as they either were not data points impacting the target variable (ID and Training_Sample) or were already captured by the target variable (Recidivism_Arrest_Year1, Recidivism_Arrest_Year2, Recidivism_Arrest_Year3). This left a total of 48 features.

Value Distributions and Significance Testing with Outcome Variable

As seen in Figure 1, there is a well-balanced distribution for the target variable with an approximate 58% belonging to the positive class.

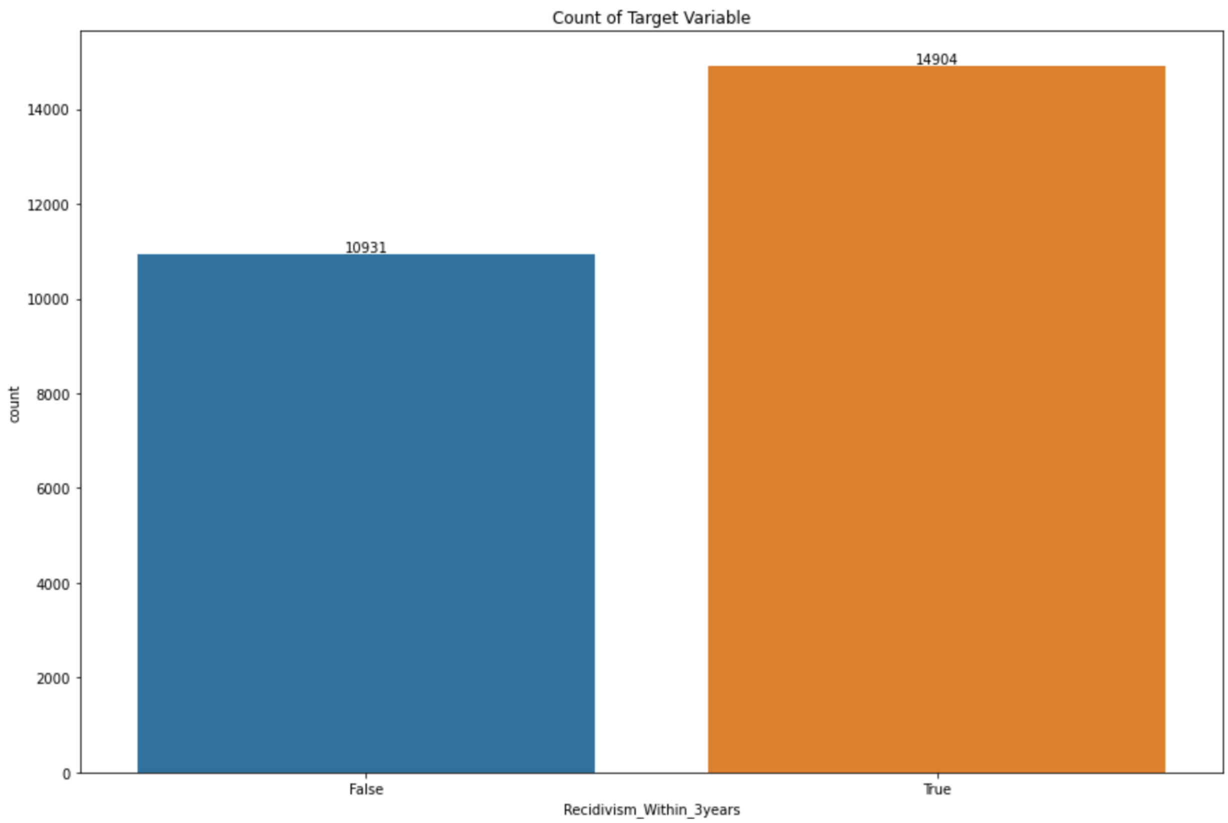


Figure 1: Count of Target Variable

Review of the features shows a highly skewed distribution for all numeric features while the majority of categorical features show a more balanced distribution. With these distributions in mind, Chi-Squared tests were performed for all categorical features while a Mann-Whitney test was performed for all numeric features to verify that a significant difference in values can be found between the positive and negative outcome.

As seen in Tables 1 and 2, all features, with the exception of Violations_ElectronicMonitoring, show a statistically significant difference between the positive and negative outcomes.

Table 1 Results of Mann-Whitney Tests with Numeric Features

	statsistic	p-value
Avg_Days_per_DrugTest	46273899.0	1.602425e-03
DrugTests_THC_Positive	59532459.0	5.699052e-82
DrugTests_Cocaine_Positive	53693536.5	2.336331e-05
DrugTests_Meth_Positive	55290700.5	7.760927e-30
DrugTests_Other_Positive	53302602.5	7.631850e-03
Percent_Days_Employed	58449431.0	2.455530e-265
Jobs_Per_Year	69129124.0	2.962256e-30

Table 2 Results of Chi-Squared Tests with Categorical Features

	chi2	p-value	dof
Gender	217.987238	2.485577e-49	1
Race	14.093703	1.739251e-04	1
Age_at_Release	826.528406	2.852097e-175	6
Residence_PUMA	114.141737	1.058508e-13	24
Gang_Affiliated	774.865907	1.573165e-170	1
Supervision_Risk_Score_First	918.863529	5.330440e-192	9
Supervision_Level_First	257.337354	1.317964e-56	2
Education_Level	380.213080	2.740177e-83	2
Dependents	50.625741	5.877916e-11	3
Prison_Offense	453.427187	7.887262e-97	4
Prison_Years	466.312893	9.520266e-101	3
Prior_Arrest_Episodes_Felony	1160.665927	4.386241e-243	10
Prior_Arrest_Episodes_Misd	835.998240	2.562639e-177	6
Prior_Arrest_Episodes_Violent	120.123184	7.259502e-26	3
Prior_Arrest_Episodes_Property	872.760150	2.090233e-186	5
Prior_Arrest_Episodes_Drug	181.546356	2.501304e-37	5
Prior_Arrest_Episodes_PPViolationCharges	1392.280727	6.477778e-299	5
Prior_Arrest_Episodes_DVCharges	112.166732	3.284902e-26	1
Prior_Arrest_Episodes_GunCharges	49.053070	2.491300e-12	1
Prior_Conviction_Episodes_Felony	288.014329	3.905084e-62	3
Prior_Conviction_Episodes_Misd	809.206850	7.781564e-174	4
Prior_Conviction_Episodes_Viol	56.108934	6.856485e-14	1
Prior_Conviction_Episodes_Prop	670.402774	5.490876e-145	3
Prior_Conviction_Episodes_Drug	112.704456	3.361541e-25	2
Prior_Conviction_Episodes_PPViolationCharges	235.790388	3.255533e-53	1
Prior_Conviction_Episodes_DomesticViolenceCharges	90.492026	1.857239e-21	1
Prior_Conviction_Episodes_GunCharges	24.337026	8.086956e-07	1
Prior_Revocations_Parole	85.352988	2.495845e-20	1
Prior_Revocations_Probation	39.147891	3.928875e-10	1
Condition_MH_SA	334.415785	1.049645e-74	1
Condition_Cog_Ed	36.186138	1.793417e-09	1
Condition_Other	4.230668	3.969980e-02	1
Violations_ElectronicMonitoring	0.345619	5.566037e-01	1
Violations_Instruction	106.636474	5.347980e-25	1
Violations_FailToReport	23.715980	1.116516e-06	1
Violations_MoveWithoutPermission	25.717312	3.952654e-07	1
Delinquency_Reports	193.103505	1.141148e-40	4
Program_Attendances	281.474974	1.271583e-54	10
Program_UnexcusedAbsences	98.323765	3.563580e-21	3
Residence_Changes	85.755185	1.786681e-18	3
Employment_Exempt	65.177668	6.844076e-16	1

As Violations_ElectronicMonitoring did not show a significant difference between the target categories it was removed from the dataset.

Correlation Heatmap

A correlation heatmap was constructed showing the strength-of-association between all features in the dataset, as well as between each feature and the target variable. Scores for continuous-continuous associations were calculated using Pearson's R while scores for continuous-categorical associations were calculated using Correlation Ratio and categorical-categorical associations were calculated using Cramer's V.

Appendix B shows the full correlation heatmap.

Missing Values

11 features contain missing values (Table 3). The missing values were dealt with as follows:

- A new category of 'Unknown' was created for the following features: Gang_Affiliated, Supervision_Level_First and Prison_Offense.
- The 4 DrugTest features were combined into one new feature, Drug_Test_Results, consisting of 3 categories ('Yes', 'No', 'Unknown') with a 'Yes' being assigned if any of the 4 DrugTest values was greater than zero. The original 4 features were then removed.
- Rows missing Percent_Days_Employed or Supervision_Risk_Score_First were removed from the dataset. Percent_Days_Employed was then binned into 4 categories: 'None', 'All', 'Less Than Half', 'More Than Half'.
- Jobs_Per_Year was found to be moderately correlated with Percent_Days_Employed and as such was removed from the dataset.
- Avg_Days_per_DrugTest showed a near zero correlation with the target variable and as such was removed from the dataset.

Table 3 Counts of Missing Values

	count
Gang_Affiliated	3167
Supervision_Risk_Score_First	475
Supervision_Level_First	1720
Prison_Offense	3277
Avg_Days_per_DrugTest	6103
DrugTests_THC_Positive	5172
DrugTests_Cocaine_Positive	5172
DrugTests_Meth_Positive	5172
DrugTests_Other_Positive	5172
Percent_Days_Employed	462
Jobs_Per_Year	808

Encoding and Scaling

As part of the training pipeline the categorical features were encoded using either One-Hot Encoding (see Appendix D for full list of encoded features) or Ordinal Encoding (see appendix C for full list of encoded features). Supervision_Risk_Score_First was scaled to a value from 0 to 1.

Train/Test Split

After the initial preprocessing steps discussed above, the dataset consisted of 24,900 observations with 42 features. The dataset was then split into separate training and testing datasets, with 20% being set aside for testing. The split was stratified on the outcome variable.

Data Modeling and Results

Models were built to predict if an individual experienced recidivism using the following methods: K-Nearest Neighbors, Logistic Regression with Elastic Net Regularization, Decision Tree, Random Forest, Support Vector Machine with Stochastic Gradient Descent and Gradient Boost.

Model Training

All models were trained using the Sci-Kit Learn library in Python[7]. The training dataset was passed through a pipeline to GridSearchCV for hyperparameter tuning. All models included the following components in their pipeline:

- OrdinalEncoder
- OneHotEncoder
- SimpleScaler
- The model being trained

After applying the Ordinal Encoder and One-Hot Encoder the datasets contained 78 features.

All models underwent 10-fold cross-validation repeated 3 times during the training phase. Specific hyperparameter tuning criteria will be discussed with each individual model. The ROC-AUC score was used during the cross-validation process to determine the optimal parameter settings. A random_state value was set as part of the cross validation definition to allow for reproducible results.

Model Evaluation Metric

Evaluation of each model trained using GridSearchCV was performed against the test dataset.

In keeping with the parameters of the National Institute of Justice's competition, the primary metric used for the final model evaluations was the Brier Score[3,9]. This allowed our results to be compared to those obtained by others.

The Brier Score measures the accuracy of binary probabilistic predictions and can be thought of as a cost function, similar to the mean squared error as utilized in regression problems. The lower the Brier Score the better the predictions are. Scores will range from 0 to 1 and are calculated as

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

where f_t is the forecasted probability, o_t is the actual outcome and N is the number of forecasted instances.

In addition to the Brier Score the precision, recall, roc_auc and F1 scores were calculated for secondary performance evaluations.

Model Details

K-Nearest Neighbors

The KNN model was trained with KNeighborsClassifier from the Sci-Kit Learn library. The tuning grid settings and resulting best parameters were as follows:

Parameter	Searched Values	Best Value
n_neighbors	[5,10, 15, 20, 50, 100, 125, 150, 200]	125
metric	['minkowski', 'chebyshev']	'minkowski'

Logistic Regression with Elastic Net Regularization

The second model was trained using LogisticRegression from the Sci-Kit Learn library. The tuning grid settings and resulting best parameters were as follows:

Parameter	Searched Values	Best Value
warm_start	[True, False]	True
penalty	['elasticnet']	'elasticnet'
C	[0.01,0.05,0.1,0.15,0.25,0.5,1.0]	0.1
l1_ratio	np.linspace(0,1,20,True)	0.736842

Decision Tree

The next model was trained using DecisionTreeClassifier from the Sci-Kit Learn library. The tuning grid settings and resulting best parameters were as follows:

Parameter	Searched Values	Best Value
criterion	['gini', 'entropy']	'entropy'
splitter	['best', 'random']	'best'

Random Forest

Next, a model was trained using RandomForestClassifier from the Sci-Kit Learn library. The tuning grid settings and resulting best parameters were as follows:

Parameter	Searched Values	Best Value
criterion	['gini', 'entropy']	'entropy'
max_depth	[5,10,15,25,50]	15

Linear SVM

The Linear SVM model was trained using SGDClassifier from the Sci-Kit Learn library. The tuning grid settings and resulting best parameters were as follows:

Parameter	Searched Values	Best Value
loss	[‘modified_huber’, ‘log’]	‘log’
penalty	[‘elasticnet’]	‘elasticnet’
alpha	[0.0001,0.001,0.01,1,10,100]	0.01
l1_ratio	np.linspace(0,1,10,True)	0

Gradient Boost

The final model was trained using XGBClassifier from the xgb library. Hyperparameter tuning was done using GridSearchCV from the Sci-Kit learn library. The tuning grid settings and resulting best parameters were as follows:

Parameter	Searched Values	Best Value
n_estimators	[1000]	1000
booster	[‘gbtree’]	‘gbtree’
eta	[0.0001,0.001,0.01,0.1]	0.01
max_depth	[3,6,8]	6
objective	[‘binary:logistic’]	‘binary:logistic’
subsample	[0.25,0.5,0.75,1]	0.75
tree_method	[‘hist’, ‘approx’]	‘approx’
eval_metric	[‘auc’]	‘auc’
min_child_weight	[1,3,5,7]	1
max_delta_step	[0,1,3,5]	1

For this model the training dataset was further divided into a training and validation set, with 20% of the original training dataset being used for the validation set. The validation set was used during training to allow for early stopping if no improvement was found in the metric score after 50 rounds.

Model Evaluations

All models were evaluated against the test dataset. Metric scores were calculated for Brier Score, F1 Score, Recall, Precision and ROC-AUC score.

Table 4 list the evaluation metric scores for each model, while Figure 2 shows the ROC curves and Figure 3 shows a confusion matrix for each model.

Table 4: Model Evaluation Metric Scores

Model	Brier Score	F1 Score	Recall	Precision	ROC_AUC
Decision Tree	0.398594	0.659110	0.657192	0.661040	0.589761
KNN	0.204805	0.759214	0.846575	0.688196	0.728433
Logistic Regression	0.198255	0.753787	0.801027	0.711808	0.747962
Random Forest	0.195874	0.765405	0.838014	0.704375	0.755605
Linear Support Vector	0.198282	0.752395	0.793493	0.715344	0.747847
Gradient Boost	0.192160	0.757820	0.804795	0.716027	0.765290

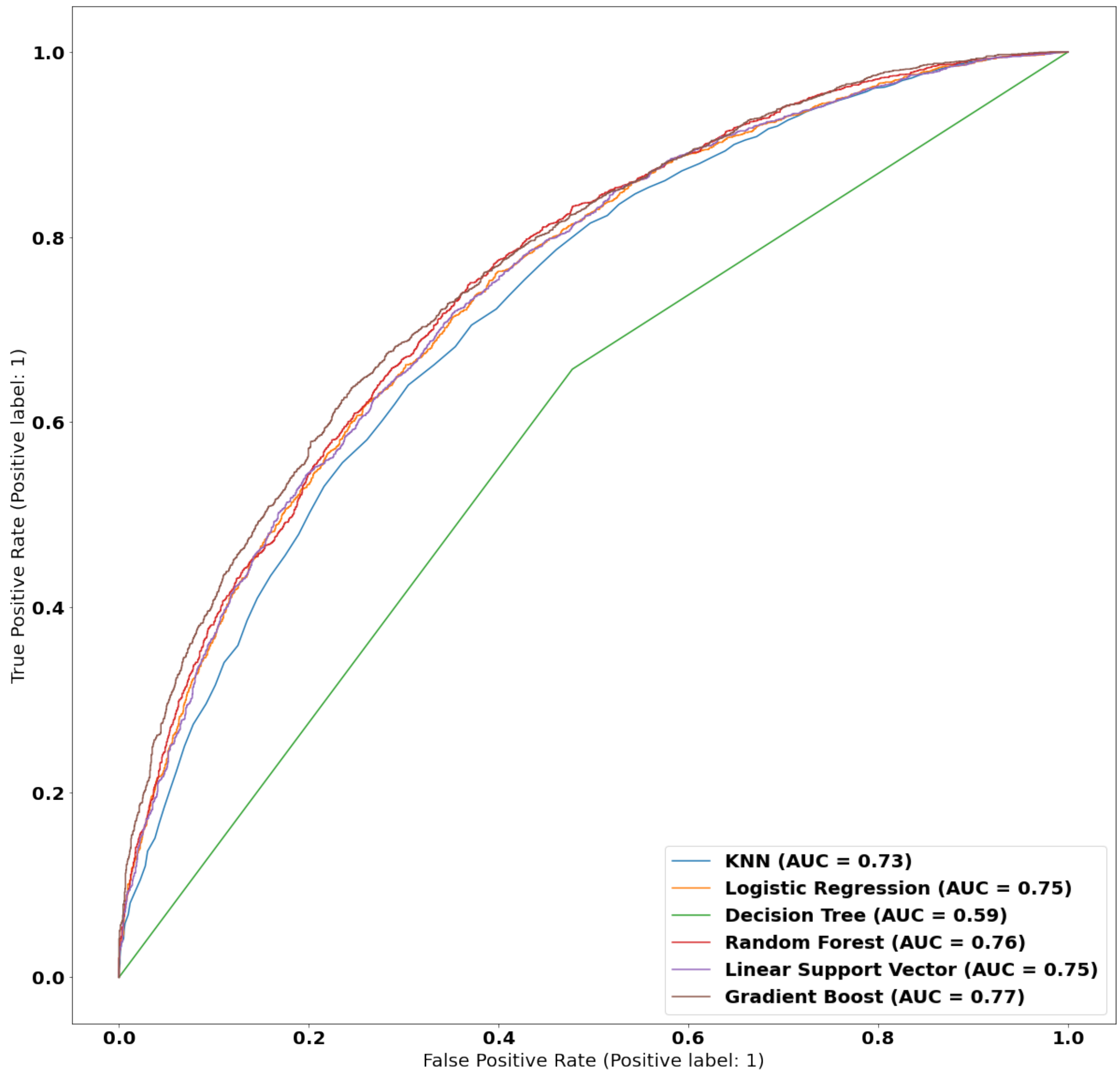


Figure 2: Model ROC Curves

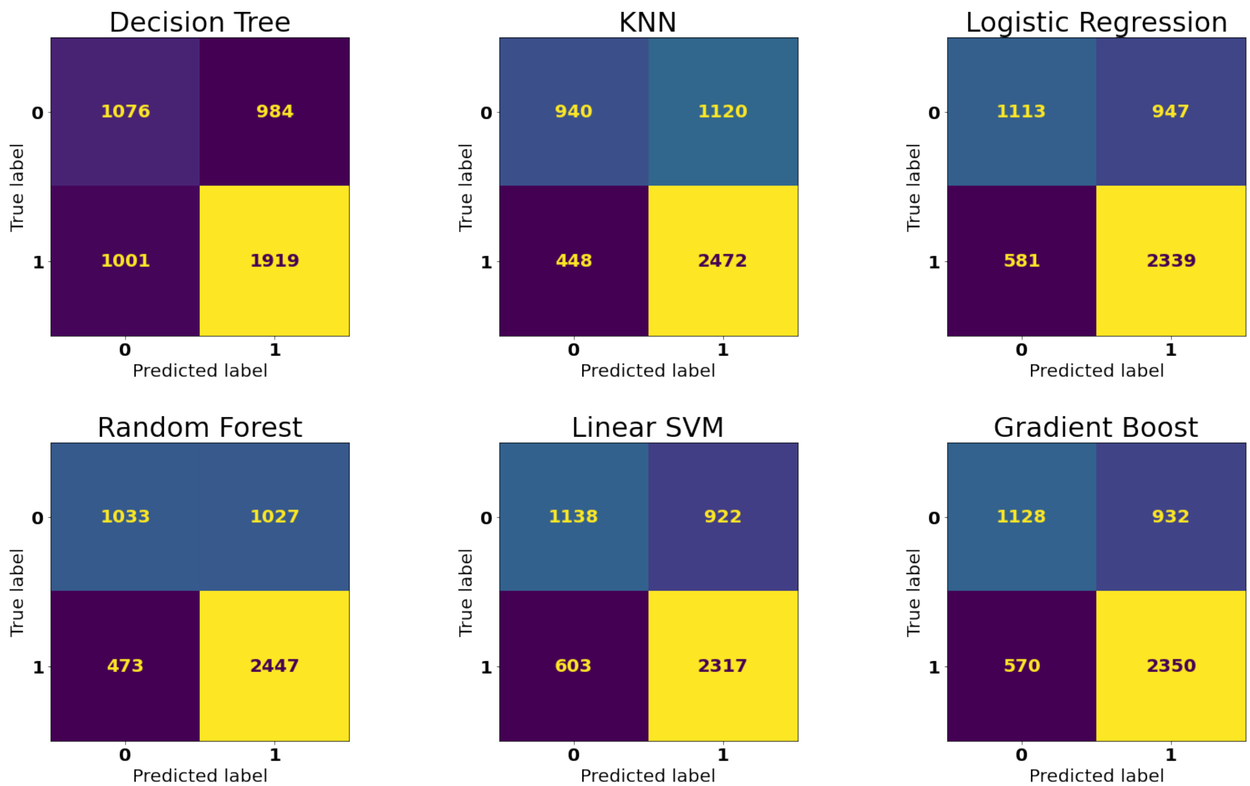


Figure 3: Model Confusion Matrices

With a Brier Score of 0.192160 the Gradient Boost model outperforms the other models. This model also has the better ROC-AUC score and Precision score.

Figure 4 shows the top 20 features in terms of feature importance.

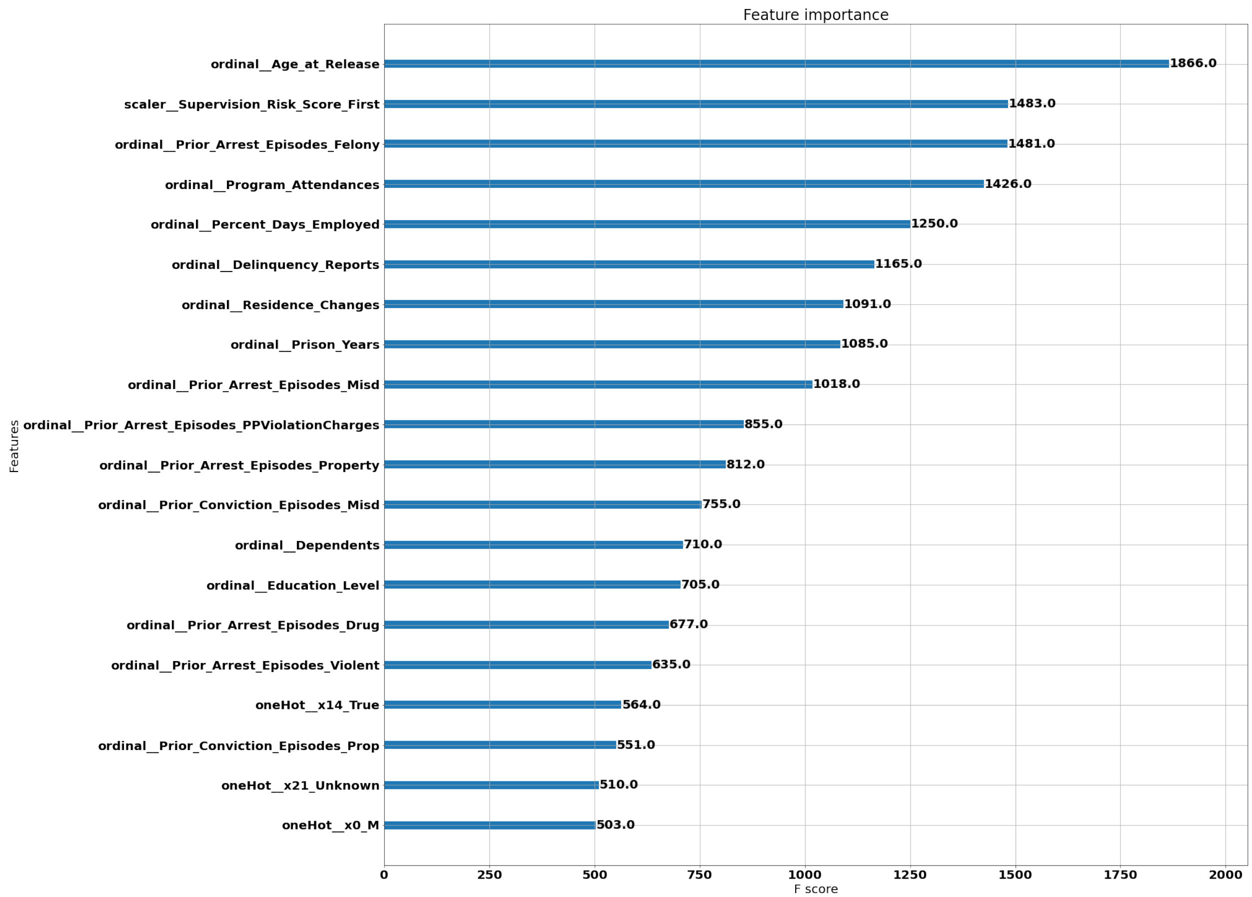


Figure 4 XGBoost Feature Importance, top 20

Conclusion

In training a model to predict an individual's likelihood of experiencing recidivism within 3 years of release from prison to parole a XGBoost model was found to provide the best performance compared to a K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest and LinearSVM classifier.

With a Brier Score of 0.192160 our model did not perform as well as the top performing models submitted to the NIJ's Recidivism Forecasting Challenge. Further hyperparameter tuning on the XGB classifier, and potentially a modified data preparation phase may see an improvement in the predictive strength of this model.

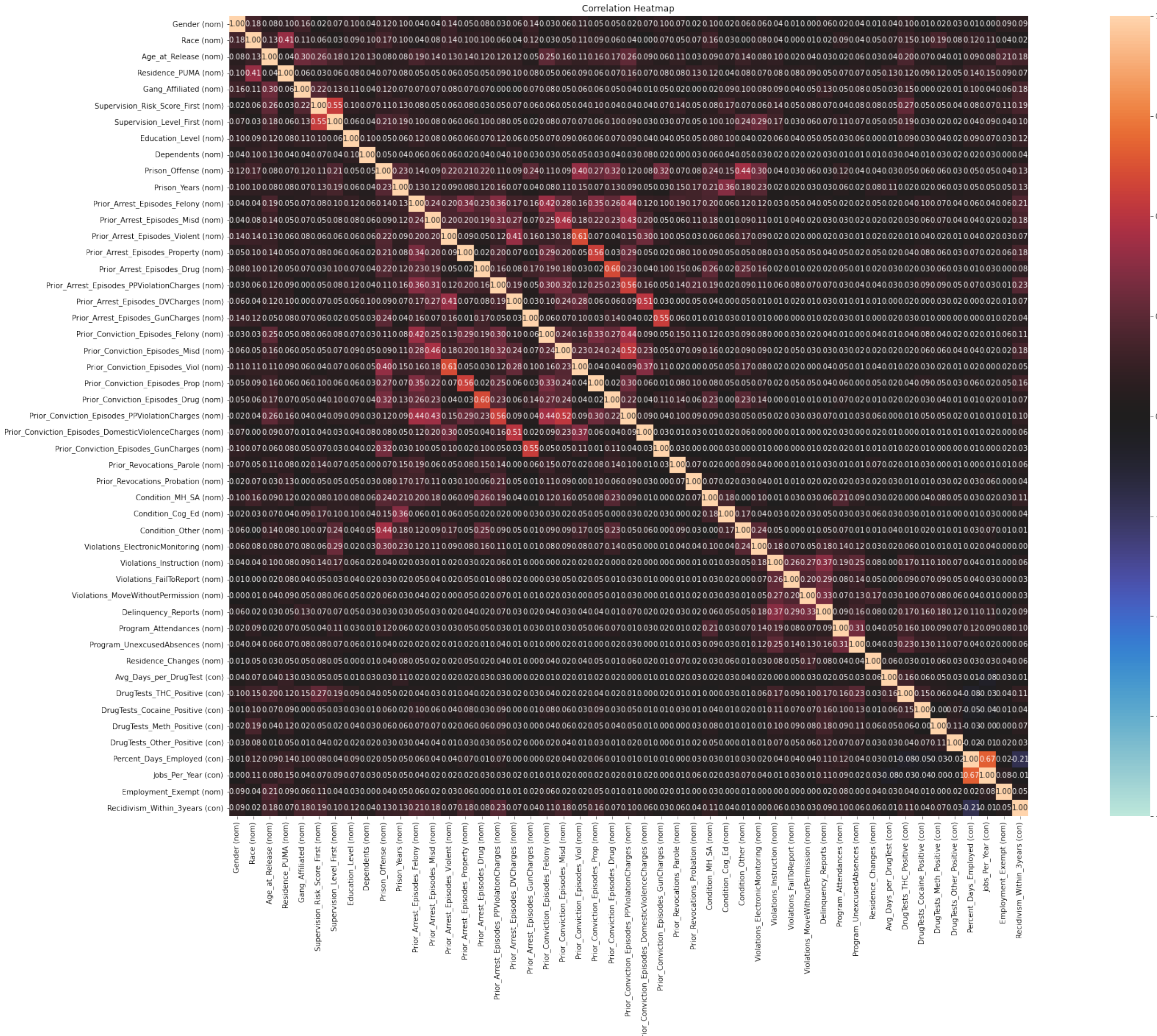
While it would be useful for the predicted probabilities to be more decisive, the ROC-AUC score of 0.77 indicates that the model performs at an acceptable level for an overall prediction of recidivism.

In addition to predicting recidivism, the XGBoost model provides insight into features/conditions that can impact the likelihood of recidivism, such as stable housing, education level and age upon release. Some conditions can't be influenced, but there are others, such as education level and employment status, that can be targeted with a goal of reducing the likelihood of recidivism.

Appendix A – Data Dictionary

Feature	Description
ID	Unique Person ID
Gender	Gender (M=Male/F=Female)
Race	Race (Black or White)
Age_at_Release	Age Group at Time of Prison Release (18-22, 23-27, 28-32, 33-37, 38-42, 43-47, 48+)
Residence_PUMA*	Residence US Census Bureau PUMA Group* at Prison Release
Gang_Affiliated	Verified by Investigation as Gang Affiliated
Supervision_Risk_Score_First	First Parole Supervision Risk Assessment Score (1-10, where 1=lowest risk)
Supervision_Level_First	First Parole Supervision Level Assignment (Standard, High, Specialized)
Education_Level	Education Grade Level at Prison Entry (<high school, High School diploma, at least some college)
Dependents	# Dependents at Prison Entry (0, 1, 2, 3+)
Prison_Offense	Primary Prison Conviction Offense Group (Violent/Sex, Violent/Non-Sex, Property, Drug, Other)
Prison_Years	Years in Prison Prior to Parole Release (<1, 1-2, 2-3, 3+)
Prior_Arrest_Episodes_Felony	# Prior GCIC Arrests with Most Serious Charge=Felony
Prior_Arrest_Episodes_Misdemeanor	# Prior GCIC Arrests with Most Serious Charge=Misdemeanor
Prior_Arrest_Episodes_Violent	# Prior GCIC Arrests with Most Serious Charge=Violent
Prior_Arrest_Episodes_Property	# Prior GCIC Arrests with Most Serious Charge=Property
Prior_Arrest_Episodes_Drug	# Prior GCIC Arrests with Most Serious Charge=Drug
Prior_Arrest_Episodes_PPViolationCharges	# Prior GCIC Arrests with Probation/Parole Violation Charges
Prior_Arrest_Episodes_DomesticViolenceCharges	Any Prior GCIC Arrests with Domestic Violence Charges
Prior_Arrest_Episodes_GunCharges	Any Prior GCIC Arrests with Gun Charges
Prior_Conviction_Episodes_Felony	# Prior GCIC Felony Convictions with Most Serious Charge=Felony
Prior_Conviction_Episodes_Misdemeanor	# Prior GCIC Convictions with Most Serious Charge=Misdemeanor
Prior_Conviction_Episodes_Violent	Any Prior GCIC Convictions with Most Serious Charge=Violent
Prior_Conviction_Episodes_Property	# Prior GCIC Convictions with Most Serious Charge=Property
Prior_Conviction_Episodes_Drug	# Prior GCIC Convictions with Most Serious Charge=Drug
Prior_Conviction_Episodes_PPViolationCharges	Any Prior GCIC Convictions with Probation/Parole Violation Charges
Prior_Conviction_Episodes_DomesticViolenceCharges	Any Prior GCIC Convictions with Domestic Violence Charges
Prior_Conviction_Episodes_GunCharges	Any Prior GCIC Convictions with Gun Charges
Prior_Revocations_Parole	Any Prior Parole Revocations
Prior_Revocations_Probation	Any Prior Probation Revocations
Condition_MH_SA	Parole Release Condition = Mental Health or Substance Abuse Programming
Condition_Cog_Ed	Parole Release Condition = Cognitive Skills or Education Programming
Condition_Other	Parole Release Condition = No Victim Contact or Electronic Monitoring or Restitution or Sex Offender Registration/Program
Violations_ElectronicMonitoring	Any Violation for Electronic Monitoring
Violations_InstructionsNotFollowed	Any Violation for Not Following Instructions
Violations_FailToReport	Any Violation for Failure to Report
Violations_MoveWithoutPermission	Any Violation for Moving Without Permission
Delinquency_Reports	# Parole Delinquency Reports
Program_Attendances	# Program Attendances
Program_UnexcusedAbsences	# Program Unexcused Absences
Residence_Changes	# Residence Changes/Moves (new zip codes) During Parole
Avg_Days_per_DrugTest	Average Days on Parole Between Drug Tests
DrugTests_THC_Positive	% Drug Tests Positive for THC/Marijuana
DrugTests_Cocaine_Positive	% Drug Tests Positive for Cocaine
DrugTests_Meth_Positive	% Drug Tests Positive for Methamphetamine
DrugTests_Other_Positive	% Drug Tests Positive for Other Drug
Percent_Days_Employed	% Days Employed While on Parole
Jobs_Per_Year	Jobs Per Year While on Parole
Employment_Exempt	Employment is Not Required (Exempted)
Recidivism_Within_3years	New Felony/Mis Crime Arrest within 3 Years of Supervision Start
Recidivism_Arrest_Year1	Recidivism Arrest Occurred in Year 1
Recidivism_Arrest_Year2	Recidivism Arrest Occurred in Year 2
Recidivism_Arrest_Year3	Recidivism Arrest Occurred in Year 3
Training_Sample	Belongs to training set

Appendix B - Correlation Matrix Heatmap



Appendix C - Features Undergoing Ordinal Encoding

Age_at_Release
Education_Level
Dependents
Prison_Years
Prior_Arrest_Episodes_Felony
Prior_Arrest_Episodes_Misd
Prior_Arrest_Episodes_Violent
Prior_Arrest_Episodes_Property
Prior_Arrest_Episodes_Drug
Prior_Arrest_Episodes_PPViolationCharges
Prior_Conviction_Episodes_Felony
Prior_Conviction_Episodes_Misd
Prior_Conviction_Episodes_Prop
Prior_Conviction_Episodes_Drug
Delinquency_Reports
Program_Attendances
Program_UnexcusedAbsences
Residence_Changes
Percent_Days_Employed

Appendix D - Features Undergoing One-Hot Encoding

Gender
Race
Residence_PUMA
Gang_Affiliated
Supervision_Level_First
Prison_Offense
Prior_Arrest_Episodes_DVCharges
Prior_Arrest_Episodes_GunCharges
Prior_Conviction_Episodes_Viol
Prior_Conviction_Episodes_PPViolationCharges
Prior_Conviction_Episodes_DomesticViolenceCharg
Prior_Conviction_Episodes_GunCharges
Prior_Revocations_Parole
Prior_Revocations_Probation
Condition_MH_SA
Condition_Cog_Ed
Condition_Other
Violations_Instruction
Violations_FailToReport
Violations_MoveWithoutPermission
Employment_Exempt
Drug_Test_Results*

**Drug_Test_Results is an engineered field using the following original fields: DrugTests_THC_Positive, DrugTests_Cocaine_Positive, DrugTests_Meth_Positive, DrugTests_Other_Positive*

Acknowledgements

I would like to thank the National Institute of Justice and the State of Georgia for their efforts in collecting and making publicly available the data used for this project.

I would also like to thank my husband, Michael D. Leavell, for his support of my educational goal over the past 2-½ years.

References

1. Atchison, N. (2018, April 20). *Community Organizations Have Important Role in Lowering Crime Rates*. Brennan Center for Justice. Retrieved October 26, 2021, from https://www.brennancenter.org/our-work/analysis-opinion/community-organizations-have-important-role-lowering-crime-rates
2. Benecchi, L. (2021, August 8). *Recidivism Imprisons American Progress*. Harvard Political Review. Retrieved October 26, 2021, from https://harvardpolitics.com/recidivism-american-progress/
3. Brier, G. (1950). VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review*, *78*(1), 1–3. https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml
4. *Incarceration Rates By Country 2021*. (2021). World Population Review. Retrieved October 26, 2021, from https://worldpopulationreview.com/country-rankings/incarceration-rates-by-country
5. *Recidivism Forecasting Challenge*. (2021). National Institute of Justice. https://nij.ojp.gov/funding/recidivism-forecasting-challenge
6. Schmitt, J., Warner, K., & Gupta, S. (2010, June). *The High Budgetary Cost of Incarceration*. Center for Economic and Policy Research. https://files.givewell.org/files/criminaljustice/The%20high%20budgetary%20costs%20of%20incarceration.pdf
7. *SciKit Learn* (1.0.1). (2021). [Machine Learning in Python]. https://scikit-learn.org/stable/index.html
8. Walmsley, R. (2021). *World Prison Population List* (twelfth edition). Institute for Criminal Policy Research. https://www.prisonstudies.org/sites/default/files/resources/downloads/wppl_12.pdf
9. Wikipedia contributors. (2021, August 23). *Brier score*. Wikipedia. https://en.wikipedia.org/wiki/Brier_score