Predicting Criminal Recidivism Using Specialized Feature Engineering and XGBoost

Article · December 2021

DOI: 10.21428/cb6ab371.d95f8c48

2 authors, including:

Suraj Rajendran
Georgia Institute of Technology
9 PUBLICATIONS **50** CITATIONS

SEE PROFILE

CrimRxiv •

Predicting Criminal Recidivism Using Specialized Feature Engineering and XGBoost

Suraj Rajendran, Prathic Sundararajan

Published on: Dec 03, 2021

Introduction

As the field of data analytics and artificial intelligence continues to evolve and expand in numerous industries, tools in these fields seem useful in helping to solve existing challenges in the criminal justice community. One of the earliest tools developed that utilized these new technologies was the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system by Northpoint[1]. Since then, a lot of analysis has been done on its performance. In addition, further tools using state of the art techniques continue to be developed in this space.

These tools are utilized to help solve various problems in the criminal justice space. One specific problem of particular interest is the existing high recidivism rates. Recidivism is defined by the US Department of Justice as “a person's relapse into criminal behavior, often after the person receives sanctions or undergoes intervention for a previous crime” [2]. It is measured by the criminal acts that result in rearrest, reconviction or return to prison in the three year period following a prisoner’s release. This metric is also often used as a way to determine the effectiveness of prisons.

As the research, development, and evaluation agency of the U.S. Department of Justice, NIJ invests in scientific research across diverse disciplines to serve the needs of the criminal justice community. In 2021, NIJ released the “Recidivism Forecasting Challenge.” With this Challenge, NIJ aims to: 1) encourage “non-criminal justice” forecasting researchers to compete against more “traditional” criminal justice forecasting researchers, building upon the current knowledge base while infusing innovative, new perspectives; and 2) compare available forecasting methods in an effort to improve person-based and place-based recidivism forecasting[3]. Our team entered into the Small Team category of the challenge and aimed to utilize state of the art machine learning techniques to assist in this field.

Methods

The aggregated dataset provided by the NIJ contains ~26,000 individuals released from Georgia prisons on discretionary parole for post-incarceration supervision between January 1st, 2013 and December 31st, 2013. NIJ split the dataset into a training and test set with a 70/30 proportion. Both the GDCS and the Georgia Bureau of Investigation provided data. The GDCS data included demographics, prison and parole case information, prior community supervision history, and supervision activities. The Georgia Bureau of Investigation provided data on prior criminal history

measures such as arrest and conviction episodes prior to prison entry. GCIC data also provides the recidivism measure, defined as a new felony or misdemeanor arrest episode within three years of parole supervision start date. This recidivism measures includes three dichotomous variables measuring if an individual recidivated in the three-year follow-up period (yes/no) as well as recidivated by time period (year 1, year 2, or year 3).

The models are being trained to predict these recidivism measures. **Supplemental Table 2** shows a list of the important features that were used for model prediction. Models predicting Year 2 and 3 recidivism were trained on supervisory data from Year 1. For Year 2 and Year 3, a feature called Early Recidivism was added. This feature delineated whether the individual already returned to prison. This could only be added to Year 2 training and Year 3 training. There would be no significance of adding it to Year 1 training since all test data would belong to the same class -- 'Not Early Recidivism'. There were multiple variables that had relatively low feature importance: Gender, Prior_Conviction_Episodes_PPViolationCharge, Prior_Conviction_Episodes_DomesticViolenceCharges, to name a few. However, because computational speed wasn't a priority for this project, all variables, regardless of statistical significance, were kept in model training.

Preprocessing and model construction were performed on Python 3.8. Preprocessing functions and ML models were imported from the Python library scikit-learn. Variables were split into categorical, ordinal, and numerical. Categorical variables were one-hot-encoded and ordinal variables were integer-encoded. Numerical variables were scaled to be between 0 and 1. Missing values were imputed using the SimpleImputer library.

An XGBoost model was used for recidivism prediction across all years. XGBoost is a supervised learning method that is based on function approximation by optimizing specific loss functions as well as applying several regularization techniques[4]. The learning method is based on the Gradient Boosted Trees algorithm. Parameters of the XGBoost were optimized using grid search hyperparameter tuning. Parameters that were optimized for are number of estimators, learning rate, and max depth of XGBoost trees. For each model trained, 10-fold validation was performed to measure average performance. Metrics captured were brier score and F1 score. Formulae for calculating brier score and F1 score are shown in **Figure 1**.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

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

Figure 1: Formulae for two performance metrics. F1 Score is shown on the left. Brier Score is shown on the right

Results

In Round 1 of the competition, we analyzed different model performances using 10 fold cross validation and found that on average XGBoost performed the best. Initially, we did hyper-parameter tuning at a broad scale for XGBoost and then we did one final tuning session with a much tighter grid of parameters. **Table 1** shows the final parameters we used for our model.

Parameter	Value
Colsample_bytree	0.75
learning_rate	0.1456
max_depth	6

min_child_weight	2
n_estimators	925
subsample	0.7

Table 1: Optimized Parameters for the XGBoost Models

In terms of results, we have brier scores from each model for the dataset. **Table X** shows our performance for each model along with the F1 score.

Model #	Brier Score	F1 Score
Year 1	0.1837	0.3009
Year 2	0.1172	0.2512
Year 3	0.0720	0.0492

Table 2: Performance Metrics for XGBoost Models across 3 Years

In terms of our performance in the challenge, we found that our Year 2 model performed very well on the hidden dataset when compared with other submissions. **Supplemental Table 3** and **Supplemental Table 4** shows the scores of our model and how they rank in comparison with other submissions to the challenge specifically for Year 2.

Discussion

The purpose of this challenge was to develop machine learning models that could accurately predict criminal recidivism across multiple years. The models trained here have taken steps toward implementation of fair models in the criminal recidivism space. While multiple model architectures were experimented with, the XGBoost architecture provided the most consistent results. The Random Forest performed well but had a much more variable performance than the XGBoost. Other regression and classification methods were assessed but performances varied widely, none achieving the success of the XGBoost. The methods of some of these results are depicted in **Supplemental Table 2**.

All models suffered from heavy class imbalance due to the relative proportion of the non-recidiviated class to the recidivated class. Synthetic augmentation methods like

SMOTE were used to mitigate this class imbalance, but did not provide significant benefits[5]. For future studies, use of a more balanced dataset or more intricate augmentation methods could aid in increasing performance of trained models.

Different probability thresholds were experimented to optimize XGBoost predictions. With the training and testing data split that was used, optimal brier scores were achieved by using a 0.5 threshold. Increasing or decreasing the threshold leads to worse metrics. Hence, the default threshold of 0.5 is ideal for achieving better performance.

From this work, important variables for predicting recidivism were explored. These variables can be further investigated to see if there is more information that can be extracted from them. While the models built suffer from class imbalance, data points with a high probability of being positive for recidivism are likely going to exhibit recidivism in reality. Hence, the models can be used as a supplement to existing technologies in the recidivism space.

Conclusions

As mentioned previously, the field of criminal justice has a lot to benefit from by utilizing state of the art machine learning techniques. We identified XGBoost to be the most effective model for our dataset and purpose. However another significant boost was gained in terms of performance simply by structuring our features based on the dataset. More specifically the addition of our Early Recidivism feature (Recidivism_Arrest_PrevYear) played a significant role in helping our model performance improve for the later rounds when compared with other submissions. This feature along with the other highlighted features in this paper should be explored further to see if they can be utilized to further improve the ability to predict recidivism.

Future Considerations

The Challenge deadlines and data were apt for the context it was proposed in. In the future, NIJ should consider exploring different topics. These topics would be interesting to investigate especially if they have limited prior research on them. One metric that could be used in future studies is the F1 score. The F1 score is the harmonic mean of the precision and recall, making it an effective score for delineating the performance of models which were trained on highly imbalanced data.

Supplemental

Year 1	Year 2	Year 3
Age_at_Release	Recidivism_Arrest_PrevYear	Recidivism_Arrest_PrevYear
Prior_Arrest_Episodes_Felony	Percent_Days_Employed	Percent_Days_Employed
Gang_Affiliated	Jobs_Per_Year	Jobs_Per_Year
Prison_Years	Age_at_Release	Age_at_Release
Prior_Arrest_Episodes_Property	Avg_Days_per_DrugTest	Avg_Days_per_DrugTest

Supplemental Table 1: Top five most important features for prediction of Recidivism by year

Models with Hyper-parameter Tuning	Brier Score		F1 Score	
	Mean	Standard Deviation	Mean	Standard Deviation
Neural Network	0.1525	0.00336	0.7850	0.0090
Random Forest	0.1087	0.000126	0.8390	0.00152
XGBoost	0.0945	0.000335	0.8559	0.00128

Supplemental Table 2: Ten Fold Validation Scores of Different Models

Place	Team Name	Brier Score
1st	**Oracle**	**0.1233**
2nd	MCHawks	0.1242
3rd	VT-ISE	0.1260
4th	DEAP	0.1263

Supplemental Table 3: Brier Score for Female Parolees in Year 2

Place	Team Name	Brier Score

1st	MCHawks	0.1405
2nd	**Oracle**	**0.1451**
3rd	VT-ISE	0.1472
4th	DEAP	0.1481

Supplemental Table 4: Brier Score for Male & Female Parolees in Year 2

References

1. Jackson, E., & Mendoza, C. (2020). Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not. *Harvard Data Science Review*, *2*(1). https://doi.org/10.1162/99608f92.1b3dadaa

2. Department of Justice. (n.d.). *NIJ's Definition of Recidivism*. National Institute of Justice. (n.d.). https://nij.ojp.gov/topics/corrections/recidivism.

3. Department of Justice. (n.d.). *Recidivism Forecasting Challenge*. National Institute of Justice. https://nij.ojp.gov/funding/recidivism-forecasting-challenge.

4. Chen, T., & Guestrin, C. (2016). Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

5. Kurniawati, Y. E. (2019). Class imbalanced Learning Menggunakan Algoritma SYNTHETIC MINORITY OVER-SAMPLING Technique – Nominal (smote-n) Pada dataset Tuberculosis anak. *Jurnal Buana Informatika*, 10(2), 134. https://doi.org/10.24002/jbi.v10i2.2441