AIT 508
Summer 2021

July 2021

Team 9

George Mason University

Analysis and Modeling of Inmate Recidivism Forecasting

Team Members:
Kevin Uruena
Ryan Vavra
Lucien Wobembong

Table of Contents

Contents

Table of Contents 2

The Project Title 3

Author’s Name(s): 3

Team 3

Header 3

1 Introduction 3

1.1 Problem 3

1.2 Objective 3

2 Data Encoding and Handling of Null Values 3

3 Data Visualization 4

4 Improving Model Accuracy 5

4.1 Removing Attributes 5

4.2 Sage Maker 5

4.3 Exhaustive Grid Search 5

5 Accuracy Results 6

6 Keywords 6

7 References 6

The Project Title: Analysis and Modeling of Inmate Recidivism Forecasting
Author’s Name(s): Ryan Vavra, Kevin Uruena, Lucien Wobembong
Team #: 9
Header:

- **Course Professor’s Name**: Dr. Eddy Zhang
- **Course Name, Number, and Section #**: AIT-580-001 (Summer 2021)
- **University Name**: George Mason University

1 Introduction

1.1 Problem

Recidivism “refers to a person's relapse into criminal behavior.” There is a high rate of recidivism among released prisoners. From those released in 2005 (401,288), there was an average of 5 arrests per prisoner (1,994,000) within a 9-year period. Forty-four percent of those released in 2005 were arrested during the first year following release.

1.2 Objective

Team 9 wants to create a model that will help justice departments determine the likelihood of an inmate recidivating within 3 years, and if an inmate will recidivate at all.

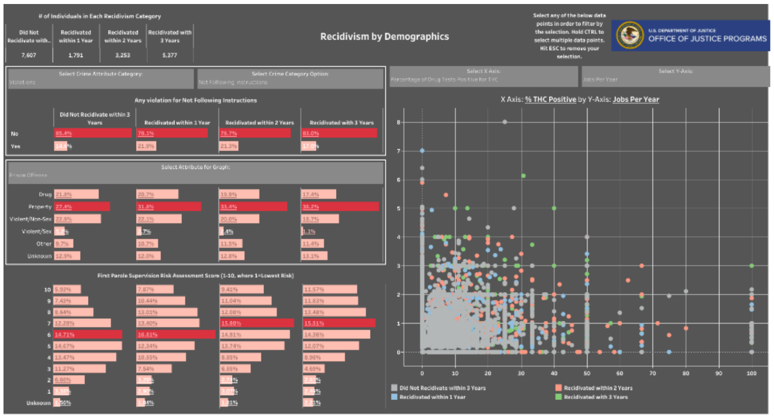
2 Data Encoding and Handling of Null Values

Team 9 decided to conduct initiative data exploration by replacing null values with zeroes for drug related attributes. We decided on this approach since we wanted to assume the principle of ‘innocent until proven otherwise.’ Replacing with mean could have added values for some inmates who did not test positive. In addition, we performed label encoding for some of the attributes. This meant converting some of the categorical values into numeric values. For example, in the original dataset, gender was either M or F, but we converted that to 0 (M) and 1 (F). Race was converted from black or white to 0 for black and 1 for white. To make prediction easier, we decided to combine “Recidivism_Within_3years,” “Recidivism_Arrest_Year1”, “Recidivism_Arrest_Year2,” and “Recidivism_Arrest_Year3.” The “Encoded Dataset” file has the complete list of all the original and their converted attributes.

Encoded Dataset

3 Data Visualization

We utilized Tableau, a data visualization software to make our dataset interactive. We chose to make it interactive so audience members who are not familiar with data analysis terminologies, methodologies or do not care for models and predictions could still understand the data and our project. The simplicity of Tableau is depicted in figure 1 and 1.1 below.

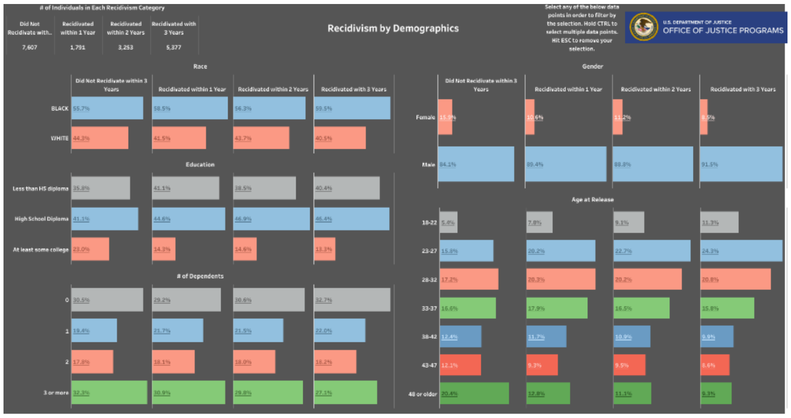


# of Individuals in Each Recidivism Category

Did Not Recidivate with..	Recidivated within 1 Year	Recidivated within 2 Years	Recidivated with 3 Years
7,607	1,791	3,253	5,377

Recidivism by Demographics
Select any of the below data points in order to filter by the selection. Hold CTRL to select multiple data points. Hit ESC to remove your selection.
U.S. DEPARTMENT OF JUSTICE
OFFICE OF JUSTICE PROGRAMS
Select Crime Attribute Category:
Violations
Select Crime Category Option:
Not Following Instructions
Any violation for Not Following Instructions
Did Not Recidivate within 3 Years
Recidivated within 1 Year
Recidivated within 2 Years
Recidivated with 3 Years
No 85.4% 78.1% 78.7% 83.0%
Yes 14.6% 21.9% 21.3% 17.0%
Select Attribute for Graph:
Prison Offense
Drug 21.8% 20.7% 19.9% 17.4%
Property 27.4% 31.8% 33.4% 38.2%
Violent/Non-Sex 22.9% 22.1% 20.0% 18.7%
Violent/Sex 5.2% 2.7% 2.4% 1.1%
Other 9.7% 10.7% 11.5% 11.4%
Unknown 12.9% 12.0% 12.8% 13.1%
First Parole Supervision Risk Assessment Score (1-10, where 1=Lowest Risk)
10 5.93% 7.87% 9.41% 11.57%
9 7.43% 10.44% 11.04% 11.83%
8 8.64% 13.01% 12.08% 13.48%
7 12.28% 13.40% 15.80% 15.51%
6 14.71% 16.81% 14.91% 14.36%
5 14.67% 12.34% 13.74% 12.07%
4 13.47% 10.55% 8.85% 8.98%
3 11.27% 7.54% 6.55% 4.69%
2 6.60% 3.29% 2.74% 2.81%
1 3.33% 2.92% 3.07% 2.49%
Unknown 1.50% 1.84% 1.81% 2.21%
Select X Axis:
Percentage of Drug Tests Positive for THC
Select Y-Axis:
Jobs Per Year
X Axis: % THC Positive by Y-Axis: Jobs Per Year
0 1 2 3 4 5 6 7 8
0 10 20 30 40 50 60 70 80 90 100
Did Not Recidivate within 3 Years
Recidivated within 1 Year
Recidivated within 2 Years
Recidivated with 3 Years

Figure 1: Recidivism by Demographic



# of Individuals in Each Recidivism Category

Did Not Recidivate with..	Recidivated within 1 Year	Recidivated within 2 Years	Recidivated with 3 Years
7,607	1,791	3,253	5,377

Recidivism by Demographics
Select any of the below data points in order to filter by the selection. Hold CTRL to select multiple data points. Hit ESC to remove your selection.
U.S. DEPARTMENT OF JUSTICE
OFFICE OF JUSTICE PROGRAMS
Race
Did Not Recidivate within 3 Years
Recidivated within 1 Year
Recidivated within 2 Years
Recidivated with 3 Years
BLACK 55.7% 58.5% 56.3% 59.5%
WHITE 44.3% 41.5% 43.7% 40.5%
Education
Less than HS diploma 35.8% 41.1% 38.5% 40.4%
High School Diploma 41.1% 44.6% 46.9% 46.4%
At least some college 23.0% 14.3% 14.6% 13.3%
# of Dependents
0 30.5% 29.2% 30.6% 32.7%
1 19.4% 21.7% 21.5% 22.0%
2 17.8% 18.1% 18.0% 18.2%
3 or more 32.3% 30.9% 29.8% 27.1%
Gender
Female 15.9% 10.6% 11.2% 8.5%
Male 84.1% 89.4% 88.8% 91.5%
Age at Release
18-22 5.4% 7.8% 9.1% 11.3%
23-27 15.8% 20.2% 22.7% 24.3%
28-32 17.2% 20.3% 20.2% 20.8%
33-37 16.6% 17.9% 16.5% 15.8%
38-42 12.4% 11.7% 10.9% 9.9%
43-47 12.1% 9.3% 9.5% 8.6%
48 or older 20.4% 12.8% 11.1% 9.3%

Figure 1.1: Recidivism by Demographic

4 Improving Model Accuracy

4.1 Removing Attributes

We went further by looking at the most important attributes in the dataset by running a random forest, which is depicted in figure 1. At first, we considered dropping any features below 0.02 (which improved our model by 0.1%), but after experimenting with various models and algorithms, we decided keeping all attributes was best for higher accuracy model. We also decided to develop two different models. The first model was indented for multiple classification, separating the data into 4 different bins corresponding to no recidivism, recidivism within 1 year, 2 years, and 3 years. The second model was indented for binary classification, with the possible values being whether an individual will or will not recidivate.

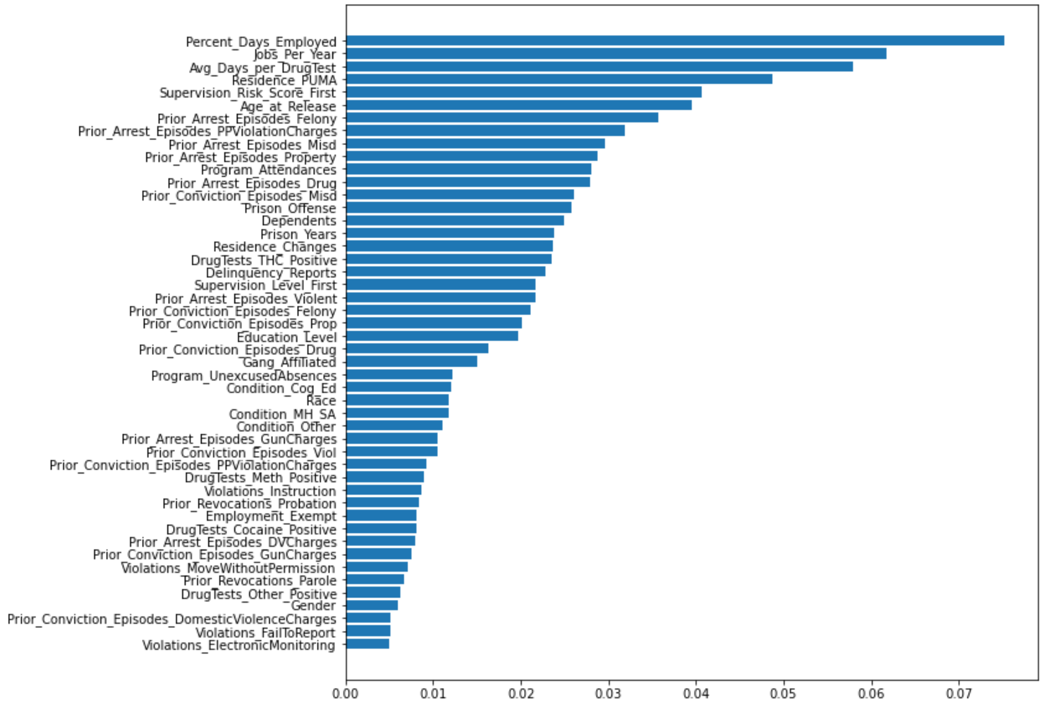


Figure 2: Features Importance

4.2 Sage Maker

Sage Maker is an AWS machine learning platform that allows the machine learning models to be easily developed and deployed. We decided to run autopilot mode in SageMaker studio to build and confirm the best performing algorithm. SageMaker selected XGBoost, trained, and tuned a model faster than our personal computers. We used the selected hyperparameters to run in local, confirming a similar performance. The model was saved in S3 and is ready for deployment in an EC2 instance.

4.3 Exhaustive Grid Search

The switch from attempting to classify four categories to trying to classify a boolean variable yielded an immediate jump in model accuracy. Despite the immediate improvement we were using only the default parameters of RF Classifier, besides n_estimators, so we hoped changing the parameters would yield an even higher accuracy score.

In addition to Sage Maker, we utilized the GridSearchCV function, imported from sklearn.model_selection, to tune our Random Forests Classifier Model. GridSearchCV is a function that allowed us to test a variety of hyperparameters and return us the set of hyperparameters that would yield the highest accuracy score.

Unfortunately, despite testing a wide variety of hyperparameter combinations, our model only improved 0.1% with the newly determined hyperparameters from the Grid Search.

5 Accuracy Results

To see which algorithms would be best for our dataset, Team 9 decided to try different algorithms and calculate the accuracy of each. As mentioned above, the 2 models developed were for binary, and multiple classification of inmate recidivism.

Algorithms	**Definition**	**Multiple Classification**	**Binary Classification**
Random Forest	Uses averaging to improve accuracy and prevent over-fitting	59.0%	73.0%
Gradient Boosting	Converts weak learners into strong learners	60.0%	74.0%
XGBoost	Like Gradient Boosting, but improved for performance	60.0%	75.0%

6 Keywords

The attached file has all the attributes and their description

NIJ Recidivism Data Codebook.xlsx

7 References

https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article

https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/

https://data.ojp.usdoj.gov/stories/s/daxx-hznc

https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html