


Predicting the Probability of Recidivism

By: Kyle Hanson, Christine Orosco, Myra Rust



Presentation Outline

- Background
- Data Cleaning and Transformation
- Preliminary Analysis
- Modeling Methods
 -  Feature Selection
 - Model Selection & Evaluation
- Model Results
- Conclusion
- Future Work

Introduction

- What is recidivism?
- 67.8% of convicts reoffend within 3 years.
- Project Objective - Use historical data and machine learning to predict the probability an individual will reoffend.
- Motivation - Provide probabilities to decision makers.




Data Cleaning & Transformation

Data Source

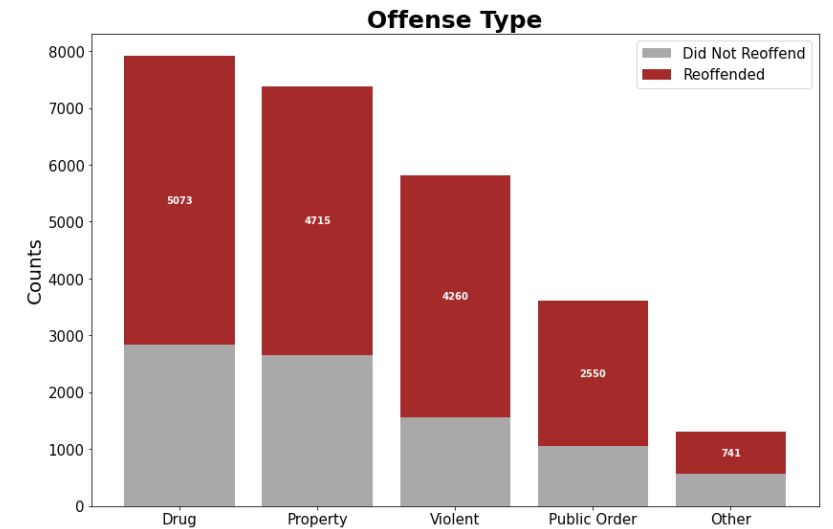
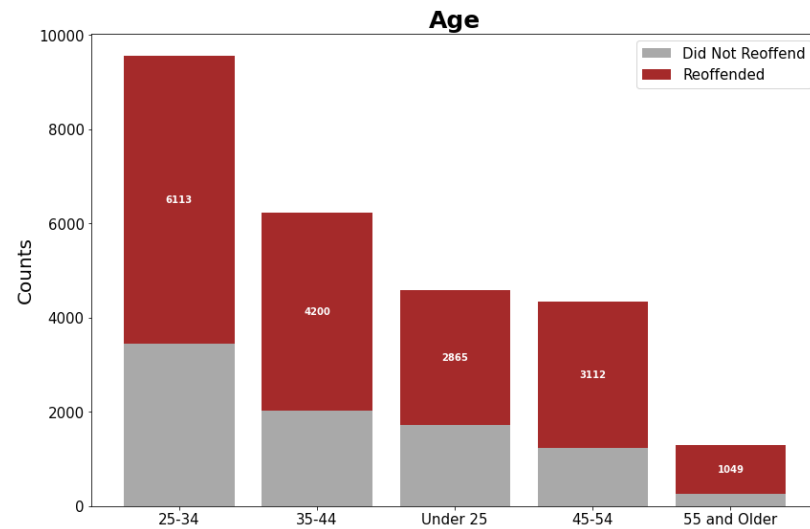
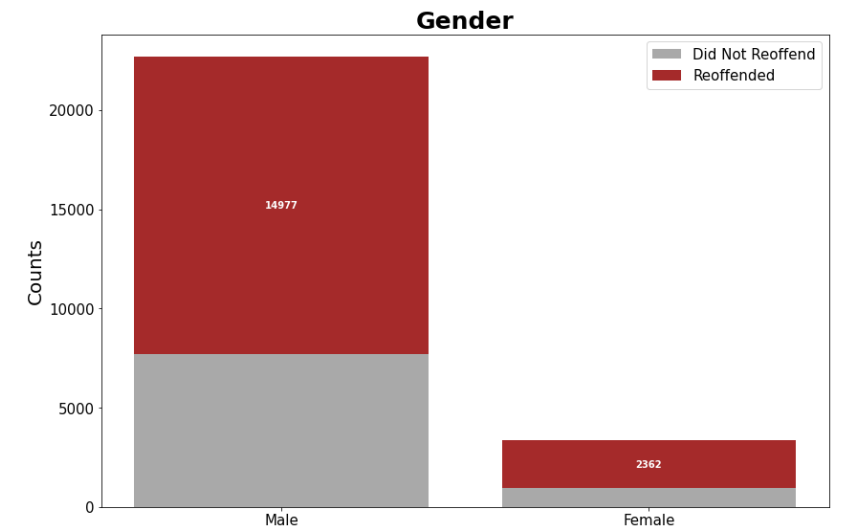
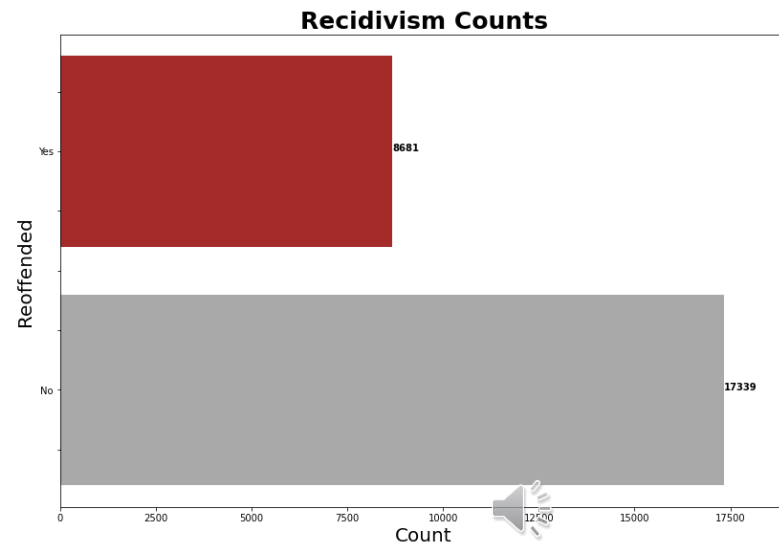
- 3-Year Recidivism for Offenders Released from Prison in Iowa (Iowa Department of Corrections, 2020).
- Data range from 2010 – 2018
- 26020 total observations

Cleaning and Transformation

- Remove variables
- Impute missing values
- Combine values
- One-hot encoded variables for model input

Features	Changes Made / Explanation
Fiscal Year Released	None
Recidivism Reporting Year	Removed – Value is always three years after value for Fiscal Year Released. 100% correlated and provided no value.
Main Supervising District	Removed – Logistical information, provided no value.
Release Type	Missing values imputed to 'Unknown'. Combined like values: <ul style="list-style-type: none">• 'Discharged – Expiration of Sentence' & 'Discharged – End of Sentence'.• 'Parole' & 'Parole Granted'.• 'Released to Special Sentence' & 'Special Sentence'.
Race – Ethnicity	Removed – Value has potential of propagating racial bias.
Age At Release 	Missing values imputed to mode '25 – 34'.
Sex	Missing values imputed to mode 'Male'.
Offense Classification	Combined like values: <ul style="list-style-type: none">• 'Other Felony (Old Code)' & 'Other Felony'
Offense Type	None
Offense Subtype	None
Return to Prison	Refactored to binary target variable. No=1, Yes=2.
Days to Return	Removed – Future data, unknown at time of prediction.
Recidivism Type	Removed – Future data, unknown at time of prediction.
New Offense Classification	Removed – Future data, unknown at time of prediction.
New Offense Type	Removed – Future data, unknown at time of prediction.
New Offense Sub Type	Removed – Future data, unknown at time of prediction.
Target Population	Removed – Logistical information, provided no value.

Exploratory Data Analysis



Correlation to Target Variable

	feature	Pearson Chi-square	p-value	Cramers V	Cramer_val
0	release	771.7566	0.0	0.1722	Strong
1	year	93.4152	0.0	0.0599	Weak
2	age	230.2147	0.0	0.0941	Weak
3	sex	28.9993	0.0	0.0334	No or very weak
4	classification	90.9506	0.0	0.0591	Weak
5	type	247.1988	0.0	0.0975	Weak
6	subtype	355.6772	0.0	0.1169	Moderate



Model Selection

- Naïve Bayes – Accuracy: 0.35%
- Logistic Regression – Accuracy: 0.59%
- Random Forest – Accuracy: 0.61%
- AdaBoost

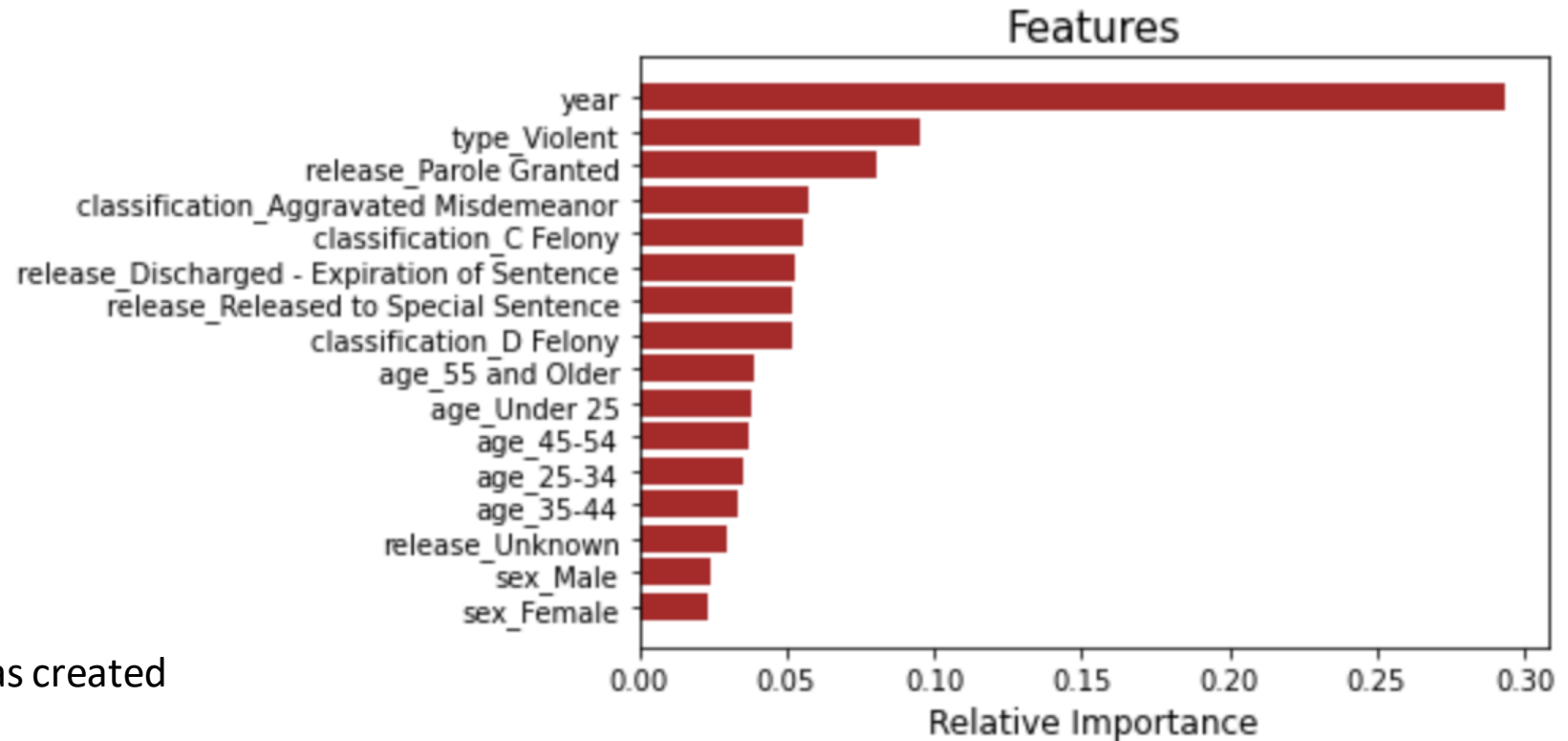
--

Random Forest complimented by AdaBoost was the most accurate.



Feature Selection

- RandomForestClassifier
- SelectFromModel
- New data frame of extracted features was created





Model Evaluation

- Split the dataset using a 70/30 train/test
- Utilized default parameters to establish a baseline metric.
- Hyperparameter tuning
- Cross validation

Random Forest Classification:

BASELINE CLASSIFICATION MODEL - 0.497

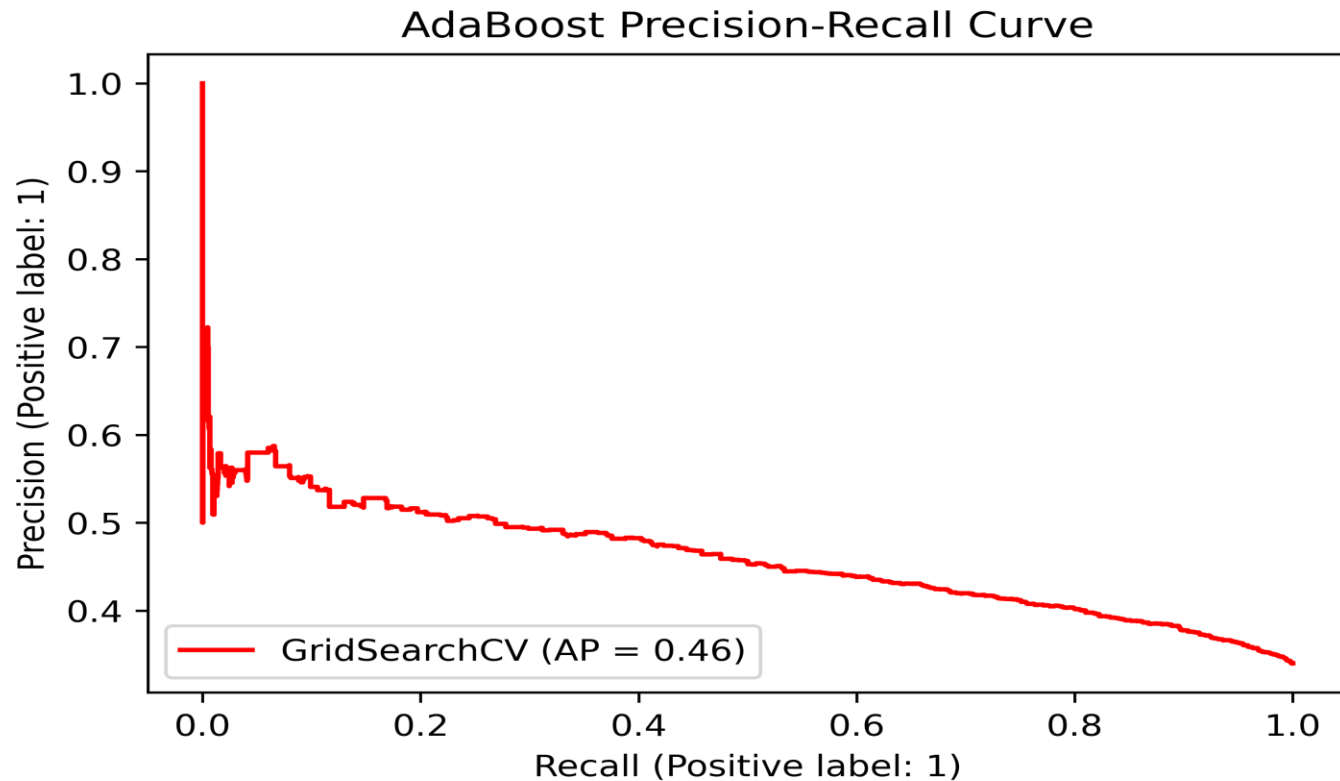
INITIAL RANDOM FOREST MODEL - 0.580

TUNED RANDOM FOREST MODEL - 0.580

ADABOOST MODEL – Mean CV Score: 0.6614

TUNED ADABOOST MODEL – Mean CV Score: 0.6624

AdaBoost Results



Mean
Accuracy

0.66

AUC

.6428

Precision

Pos 0.44

Neg 0.75

Recall

Pos 0.61

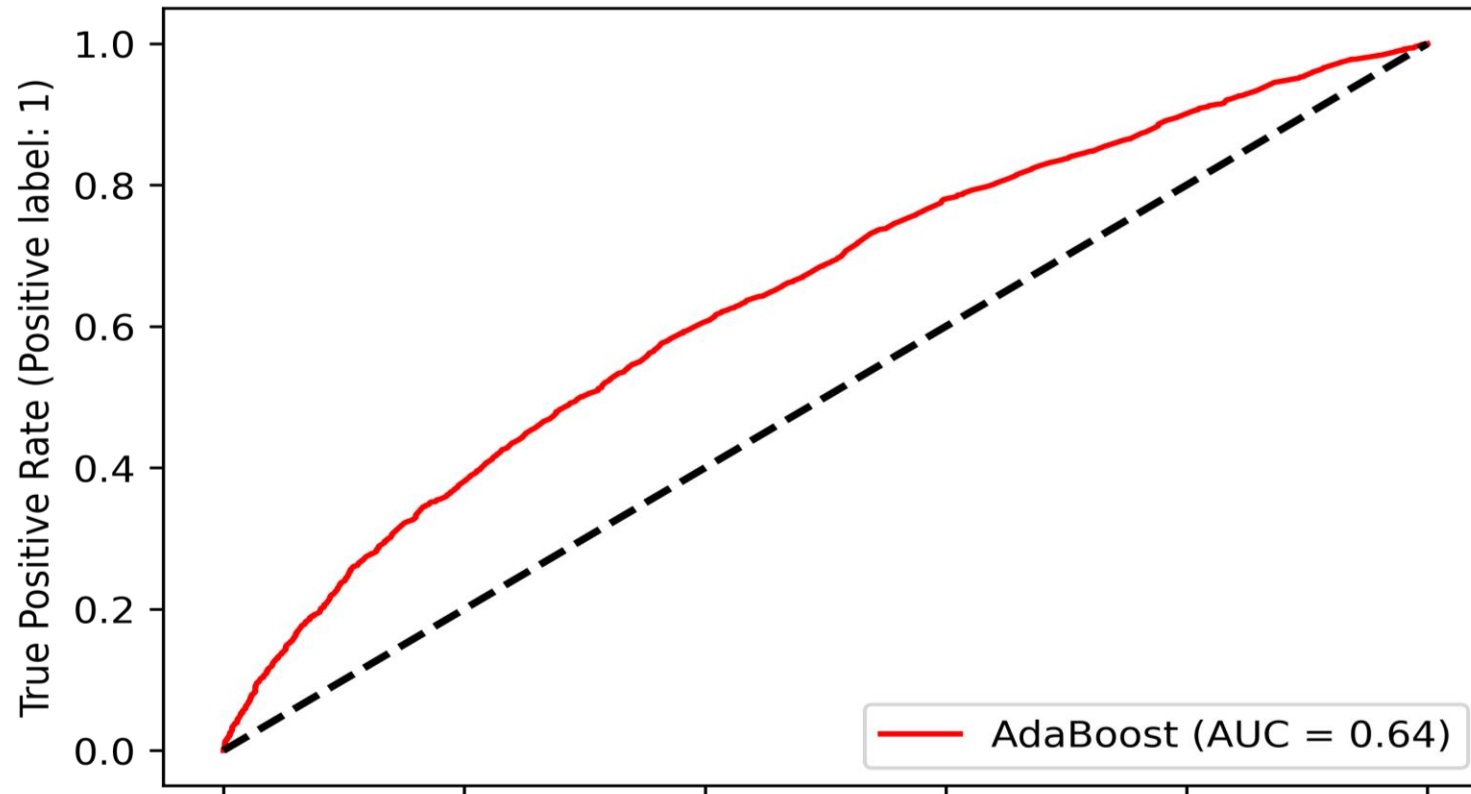
Neg 0.59

F1

Pos 0.51

Neg 0.66

AdaBoost ROC Curve



- ROC Curve indicative of poor results.
- Curve close to the zero line.
- AUC generalizes the accuracy.

Additional Models

Stacked Ensemble

- Random Forest
- Naïve Bayes
- Logistics Regression

Accuracy

- 63

Weighted Average Ensemble

- Random Forest
- Naïve Bayes
- AdaBoost

Accuracy

- 66

Neural Network

- 3 Layers
- Compile with
 - Adam Optimizer
 - Loss binary cross entropy

Accuracy

- 66



Conclusions

- Model not accurate enough for critical decisions
 - Weak or no correlation amongst variables
 - Not enough variables in the dataset
- Need additional variables to train the model
- Examples of additional variables (Florida Department of Corrections, 2020)
 - Social Cognitive factors
 - Inmate Education
 - Inmate Incarceration



Future Work

- Acquire datasets from the Florida Department of Corrections
- Explore Ensemble machine learning models
- Explore Feed Forward Neural Network model

References / Citations

Durose, M., Cooper, A., Snyder, H. (2014, April 22). Recidivism of Prisoners released in 30 states in 2005: Patterns From 2005 – 2010 – Update. Bureau of Justice Statistics. Retrieved from <https://www.bjs.gov/index.cfm?ty=pbdetail&iid=4986>.

Iowa Department of Correction. (2020, November 16). 3-Year Recidivism for Offenders Released from Prison in Iowa. Iowa Data. Retrieved from <https://data.iowa.gov/Correctional-System/3-Year-Recidivism-for-Offenders-Released-from-Pris/mw8r-vqy4>.

Florida Department of Corrections. (2020, June). Florida Prison Recidivism Report: Releases from 2008 to 2018. http://www.dc.state.fl.us/pub/recidivism/2019-2020/FDC_AR2019-20.pdf