Predicting Probability of Recidivism for Current Texas Inmates

Is an inmate’s initial crime and their age at the time it was committed predictive of recidivism?

Jacqueline Walker - Aug 2019

Why do we want to know?

This work is done in an attempt to support any entity that provides services to previously incarcerated individuals.

Recidivism

The tendency of a convicted criminal to reoffend.

The Process:

1. Research
2. Data Collection
3. Cleaning, Feature Engineering, & EDA
4. Model Fittings & Metrics
5. Model Selection & Application
6. Summary Statistics

The Process:

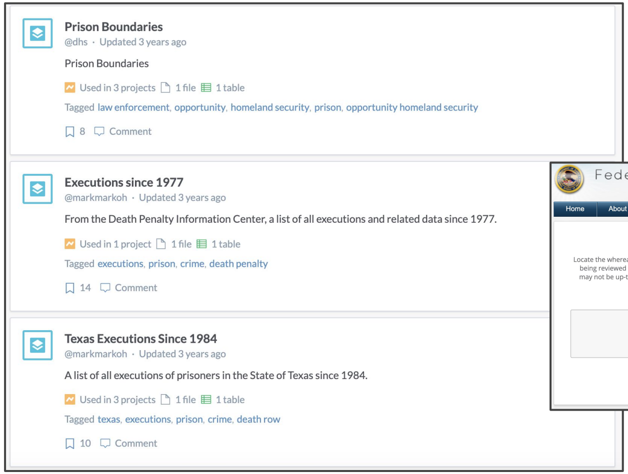
1. Research
2. Data Collection
3. Cleaning, Feature Engineering, & EDA
4. Model Fittings & Metrics
5. Model Selection & Application
6. Summary Statistics

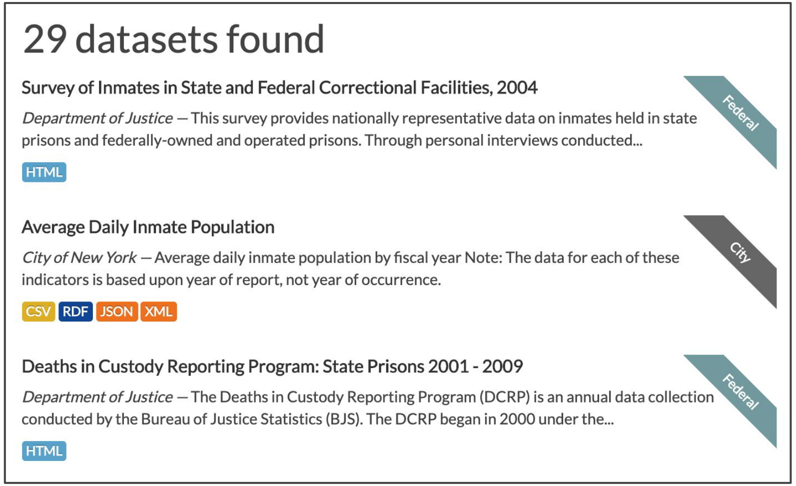
Research

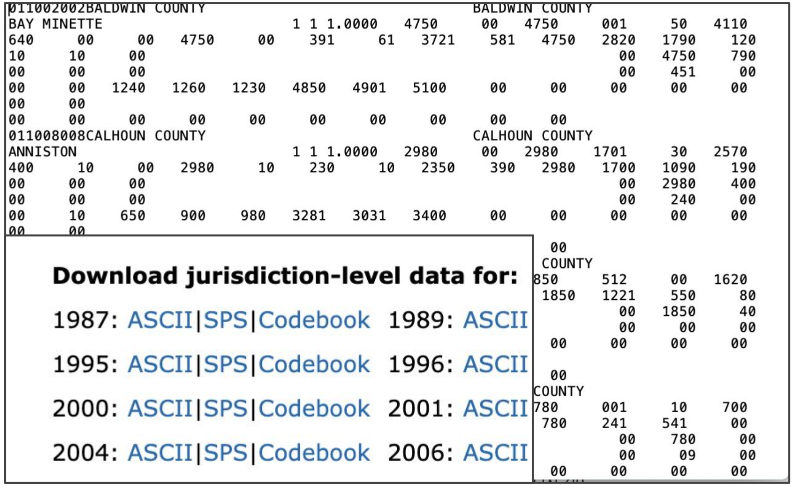
Research

---

1. Available government datasets.







Research

---

2. The COMPAS algorithm, and ProPublica’s work in exposing bias.

Data Kernels Discussion (2) Activity Metadata Download (3 MB) New Notebook

Original Article: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Original data from ProPublica: https://github.com/propublica/compas-analysis

Additional "simple" subset provided by FairML, based on the proPublica data:

http://blog.fastforwardlabs.com/2017/03/09/fairml-auditing-black-box-predictive-models.html

Inspiration

Ideas:

- Feature importance when predicting the COMPASS score itself, or recidivism/crime risks.
- Reweighting data to compensate for bias, e.g. subsetting for the violent offenders, or adjusting better for base risk.
- Feature selection based on "legal usage"/fairness (E.g. exclude race and see how well your model works. It worked for me).

Data (3 MB)

Data Sources		About this file	Columns
compas-scores-ra...	60.8k x 28	Raw Compass Score	# Person_ID
cox-violent-parsed....	18.3k x 52		# AssessmentID
cox-violent-parsed_...	18.3k x 40		# Case_ID
propublicaCompassRecidivism_da...			A Agency_Text
			A LastName
			A FirstName
			A MiddleName

So ProPublica did its own analysis.

How We Acquired the Data

We chose to examine the COMPAS algorithm because it is one of the most popular scores used nationwide and is increasingly being used in pretrial and sentencing, the so-called “front-end” of the criminal justice system. We chose Broward County because it is a large jurisdiction using the COMPAS tool in pretrial release decisions and Florida has strong open-records laws.

Through a public records request, ProPublica obtained two years worth of COMPAS scores from the Broward County Sheriff’s Office in Florida. We received data for all 18,610 people who were scored in 2013 and 2014.

Because Broward County primarily uses the score to determine whether to release or detain a defendant before his or her trial, we discarded scores that were assessed at parole, probation or other stages in the criminal justice system. That left us with 11,757 people who were assessed at the pretrial stage.

Compas Analysis

What follows are the calculations performed for ProPublica's analysis of the COMPAS Recidivism Risk Scores. It might be helpful to open the methodology in another tab to understand the following.

Loading the Data

We select fields for severity of charge, number of priors, demographics, age, sex, compas scores, and whether each person was accused of a crime within two years.

In [1]:
```
# filter dplyr warnings
%load_ext rpy2.ipython
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
%%R
library(dplyr)
library(ggplot2)
raw_data <- read.csv("./compas-scores-two-years.csv")
nrow(raw_data)
```

```
[1] 7214
```

Research

---

3. The Huntsville State Prison

THE TEXAS TRIBUNE

TEXAS PRISON INMATES > PRISON UNITS

Huntsville Unit

INMATES

Inmates

Name	Age	Main Crime	Entered On	Term	Crime Location	Home County
[redacted]	47	MURDER	6/5/2017	Life	Bexar	Harris
	51	AGG ROBBERY W/DDLY WPN	5/10/2005	Life	Dallas	Dallas
	69	MURDER	4/16/2015	Life	Angelina	Angelina
	58	BURG OF HABIT	7/8/2013	Life	Travis	Travis
	71	MURDER	11/24/1993	Life	Harris	Harris
	57	AGG ASLT AGAINST PUB SERV	6/27/2019	Life	Harris	Harris
	57	AGG ROBBERY W/DEADLY WPN	9/13/2013	Life	Dallas	Dallas
	54	MURDER	2/14/2019	Life	Nueces	Nueces
	71	ROBB BY ASLT	9/19/2006	Life	Harris	Harris
	48	ENGAGE-ORGAN CRIM ACT	2/18/2012	Life	Williamson	Dallas
	49	ATT CAPITAL MURDER	8/9/1995	Life	Bell	Bell
	42	CAPITAL MURDER	8/26/1999	Life	Tarrant	Dallas

Research

---

4. The Texas Tribune

THE TEXAS TRIBUNE

DATA APPLICATIONS

Texas Prison Inmates

Use this app to explore Texas' prison units, and learn more about the more than 137,000 inmates housed inside them. Search for inmates by name, look at the most common crimes among inmates, take a closer look at specific prison units. You can also view all Death Row inmates.

108 Units

137,949 Inmates

INMATES

THE TEXAS TRIBUNE

TEXAS PRISON INMATES

Prison Units

INMATES

Name	Type	Prisoners	Operator
Allred	Prison	3677	Correctional Institutions Division
Beto	Prison	3315	Correctional Institutions Division
Boyd	Prison	1332	Correctional Institutions Division
Bradshaw	State Jail	1849	Corrections Corporation of America
Bridgeport	Prison	520	Global Expertise in Outsourcing
Briscoe	Prison	1369	Correctional Institutions Division
Byrd	Prison	1016	Correctional Institutions Division
Chase Field Wilderness	Work Program	383	Correctional Institutions Division
Clemens	Prison	1137	Correctional Institutions Division
Clements	Prison	3738	Correctional Institutions Division
Cleveland	Prison	513	Global Expertise in Outsourcing
Coffield	Prison	4127	Correctional Institutions Division
Cole	State Jail	809	Correctional Institutions Division

Data Collection

Web Scraping:

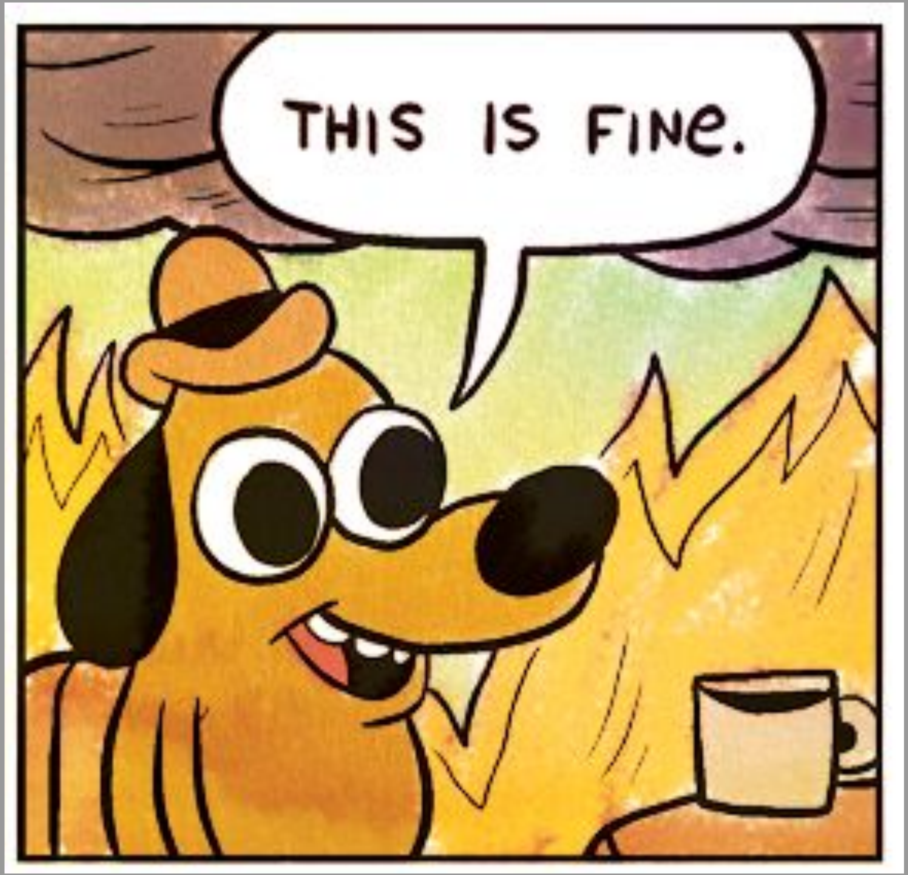
1. Beautiful Soup
2. Amazon Web Services


THIS IS FINE.

Oh, you didn't understand how it worked and you terminated your session, losing all your functions...

*Oh, you forgot useful print statements, and you're lost in the abyss of the * …*

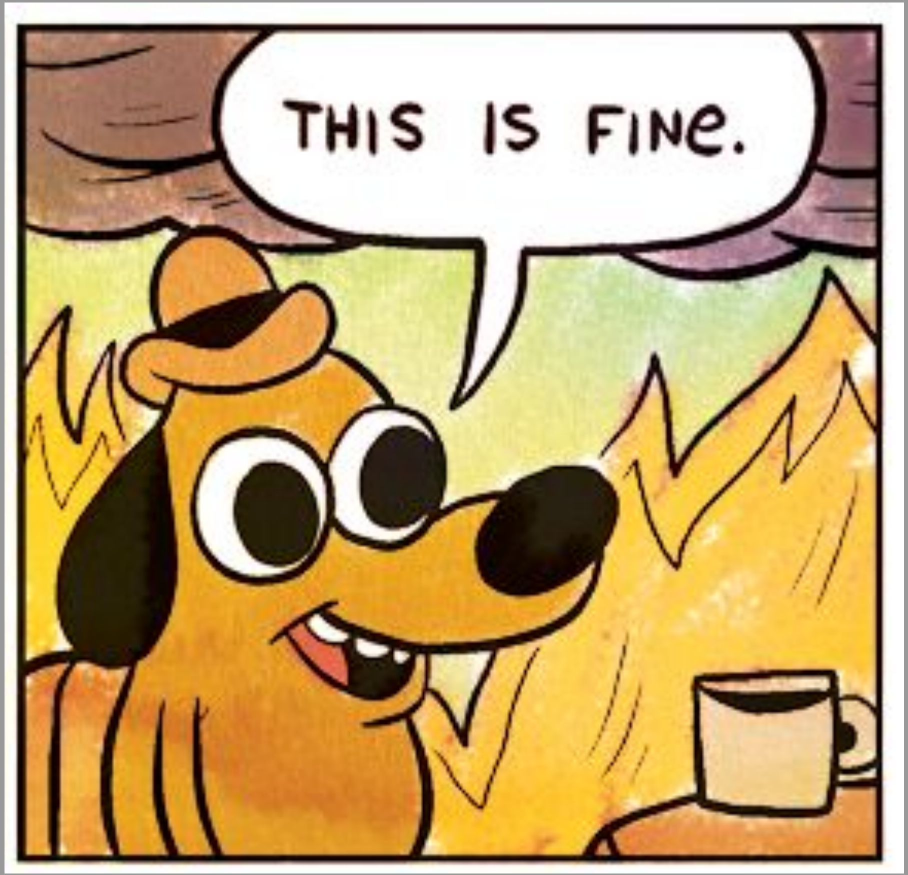
Oh, you didn't embed saving to a .csv within the function, so you have to wait until the entire thing is finished before you can do any work…



Oh, the website updated, adding a prison and reducing the number of inmates by over 1000 during your scrape…

Oh, you didn't incorporate try and except statements for every possible place data could be missing…

Oh, you didn't get overlapping information to validate the merge for each dataset you were scraping…



What was collected

[redacted]

TDCJ Criminal History

Crime	Committed On	County	Sentence	Sentence Began
CAP MURDER CHILD U/10YRS	11/19/2014	Dallas	Life Without Parole	11/21/2014
BURG HABIT	5/19/2008	Tarrant	2 years	3/29/2011

Maximum sentence	7000 years	**Sex**	Male
Projected release date	1/1/5555	**Age**	30
TDCJ ID	02053637	**Race**	Hispanic
Unit	Allred	**Height**	6 ft 2 in
DOB	2/3/1989	**Weight**	290 lbs
Home County	Dallas	**Hair Color**	Brown
		Eye Color	Brown

- Name, ***TDCJ ID***
- The current offense and three most recent priors
- Dates, term lengths, and crimes for each
- Home county, prison unit, DOB, age, race, sex, projected release date

Limitations of the data

- Only 4 most recent crimes committed per inmate, rather than all crimes for each.

- Only current inmates, not former inmates that have not reoffended.

- Multiple types of facilities, not exclusively prisons.

Prison
State Jail
Prison
Prison
Prison
Work Program

- No additional inmate information (home zipcode, occupation, family, etc.)
  - *ethical considerations with this additional information

Data Cleaning, Feature Engineering, & EDA

Cleaning, Feature Engineering & EDA

1. Merging the datasets (on a ***unique identifier*** - TDCJ ID)

2. Removing nulls for relevant columns

3. Finding the first crime committed, and the information connected with this crime for each inmate

```
def find_crime(row):
    if row['pr_crime_3'] != 'No_data':
        return row['pr_crime_3']

    elif row['pr_crime_2'] != 'No_data':

        return row['pr_crime_2']

    elif row['pr_crime_1'] != 'No_data':
        return row['pr_crime_1']

    else:
        return row['pr_crime_0']
```

```
df['feature_crime'] = df.apply(find_crime, axis=1)
```

Cleaning, Feature Engineering & EDA

1. Calculating the age of each person at the time of their *first crime.

feature_term	commit_age	feature_commit_date	DOB
60 years	32.003395	1980-03-03	1948-03-02
60 years	27.294195	1984-08-06	1957-04-21
55 years	16.895624	1994-02-02	1977-03-12
55 years	26.100467	1982-11-28	1956-10-22
55 years	23.359823	1991-09-22	1968-05-13
55 years	54.410426	1994-04-10	1939-11-12

commit_age = feature_commit_date - DOB

Cleaning, Feature Engineering & EDA

---

1. Converting term lengths of feature crime to floats.

2. Filtered out any observations where the projected release date was more than 30 years from now.

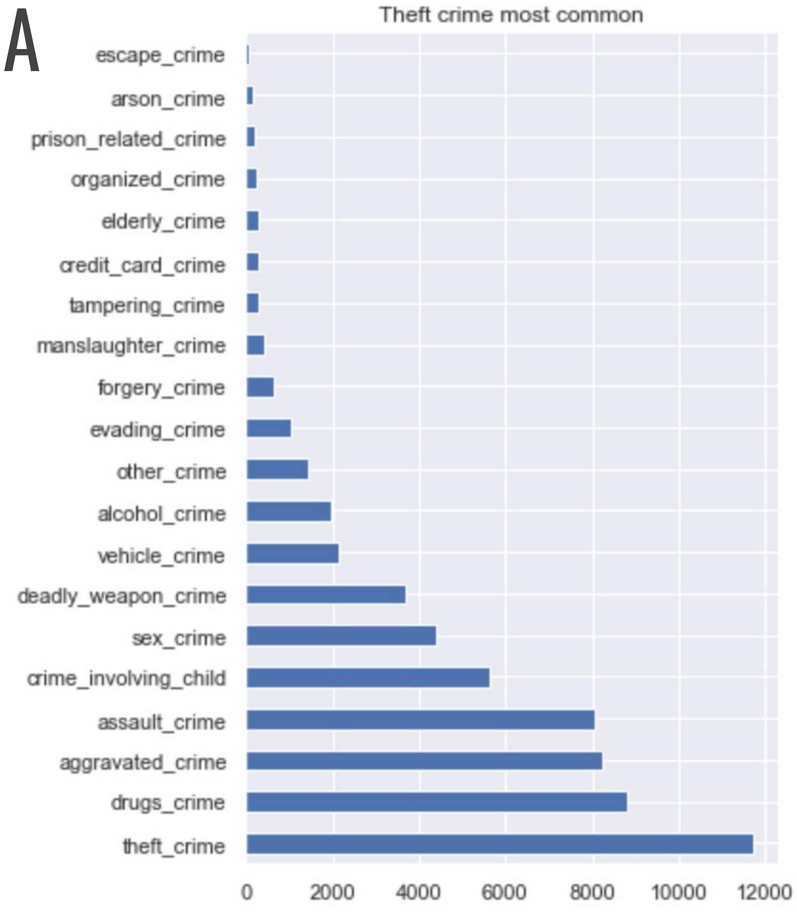
feature_term_flt	feature_term
2.000000	2 years
30.000000	30 years
8.000000	8 years
2.000000	2 years
30.000000	30 years
8.000000	8 years
0.500000	6 months
1.500000	1 year, 6 months
1.416667	1 year, 5 months
25.000000	25 years
30.000000	30 years
1.500000	1 year, 6 months

Cleaning, Feature Engineering & EDA

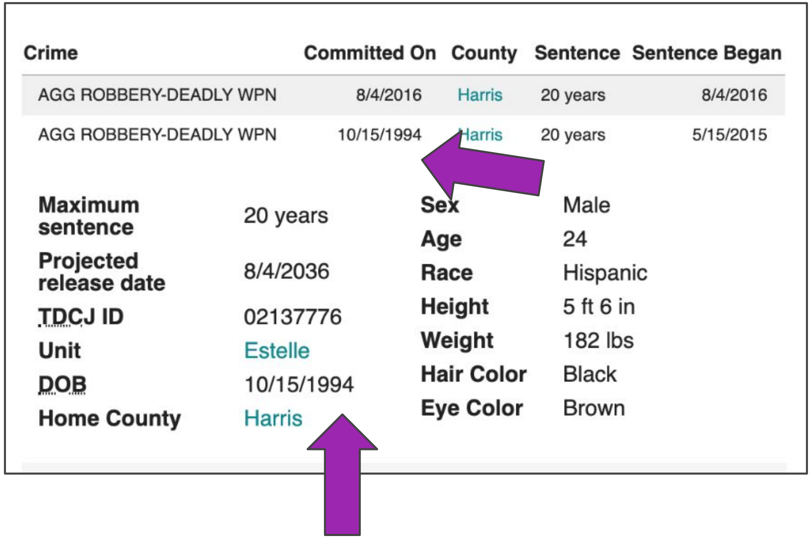
---

1. Categorical columns for types of crimes

2. Binned age and length of term, also kept as floats for comparison

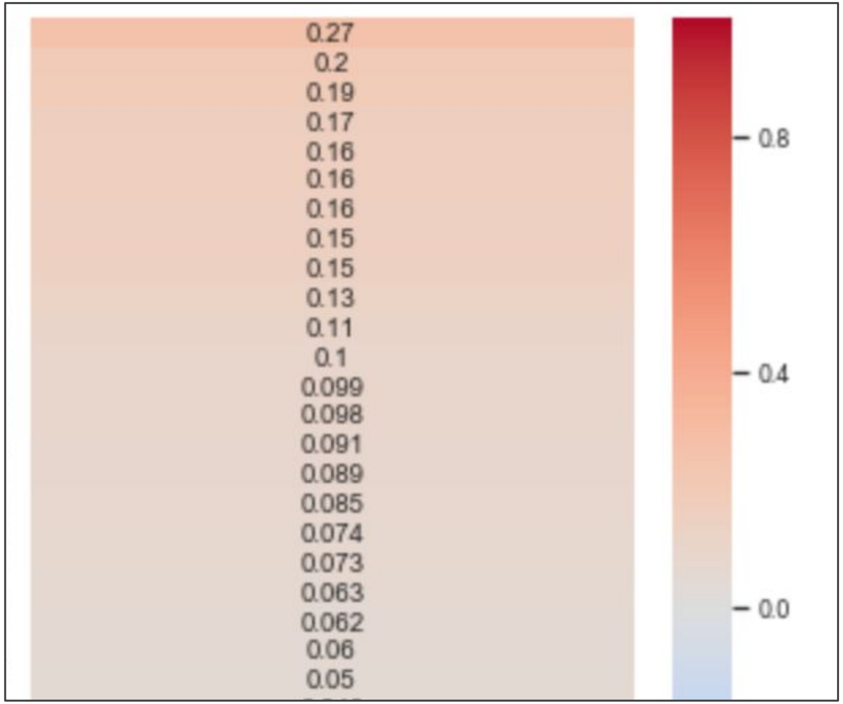


Challenges Discovered



Crime	Committed On	County	Sentence	Sentence Began
AGG ROBBERY-DEADLY WPN	8/4/2016	Harris	20 years	8/4/2016
AGG ROBBERY-DEADLY WPN	10/15/1994	Harris	20 years	5/15/2015

Maximum sentence	20 years	Sex	Male
		Age	24
Projected release date	8/4/2036	Race	Hispanic
TDCJ ID	02137776	Height	5 ft 6 in
Unit	Estelle	Weight	182 lbs
DOB	10/15/1994	Hair Color	Black
Home County	Harris	Eye Color	Brown



- .27 highest correlation with target.
- Only 12 features above .10 correlation.
- Incorrect data on the website.
- Target imbalanced classes.

```
1  y.value_counts(normalize=True)

1    0.745561
0    0.254439
Name: final_target, dtype: float64
```

Model Selection

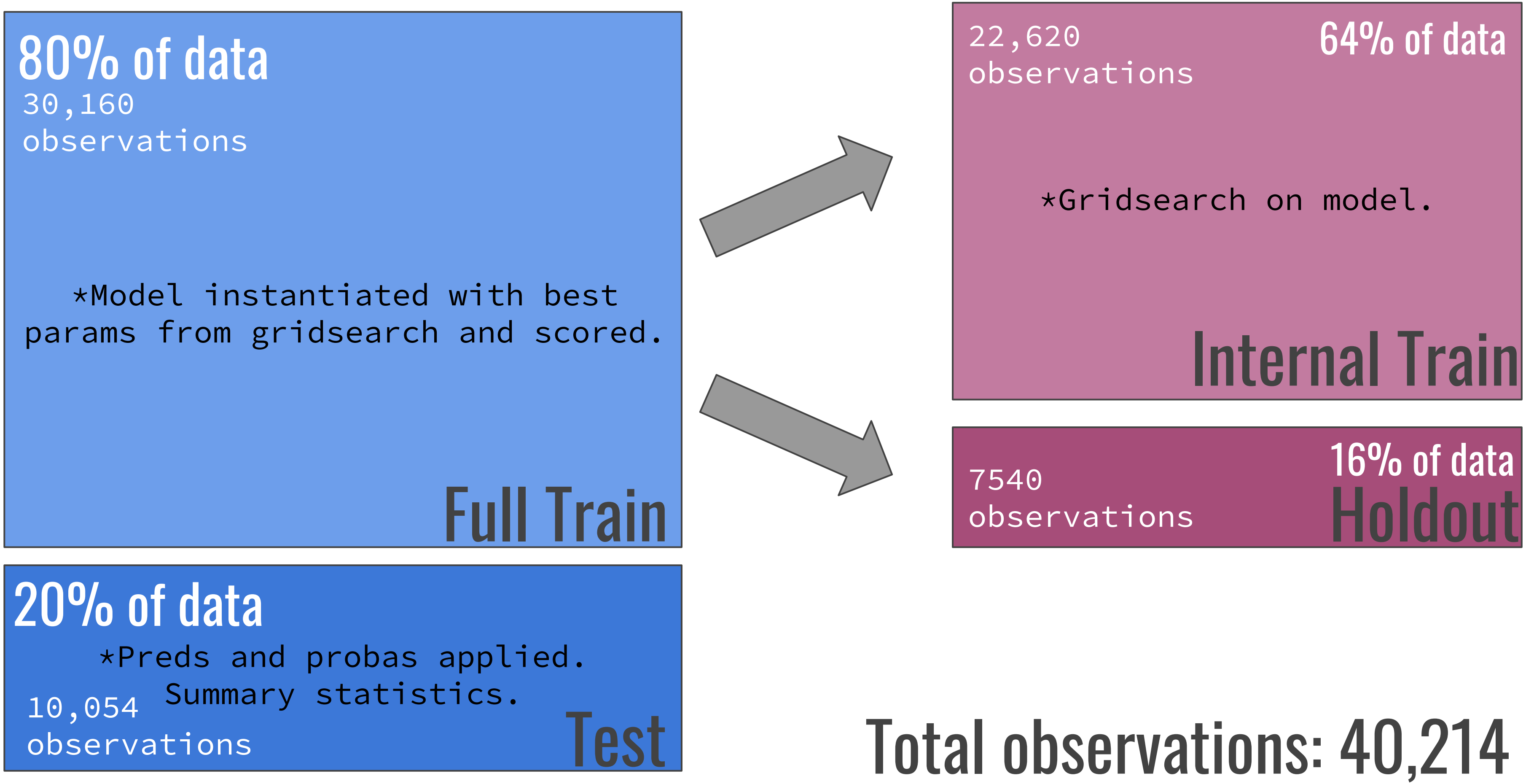
Classification Models:

1. Support Vector Classifier
2. Random Forest
3. Logistic Regression

Research for model selection in consideration of imbalanced classes led to the selection of SVC and random forest. I also wanted to do logistic regression to have access to the coefficients.

*research article link

For Each Model

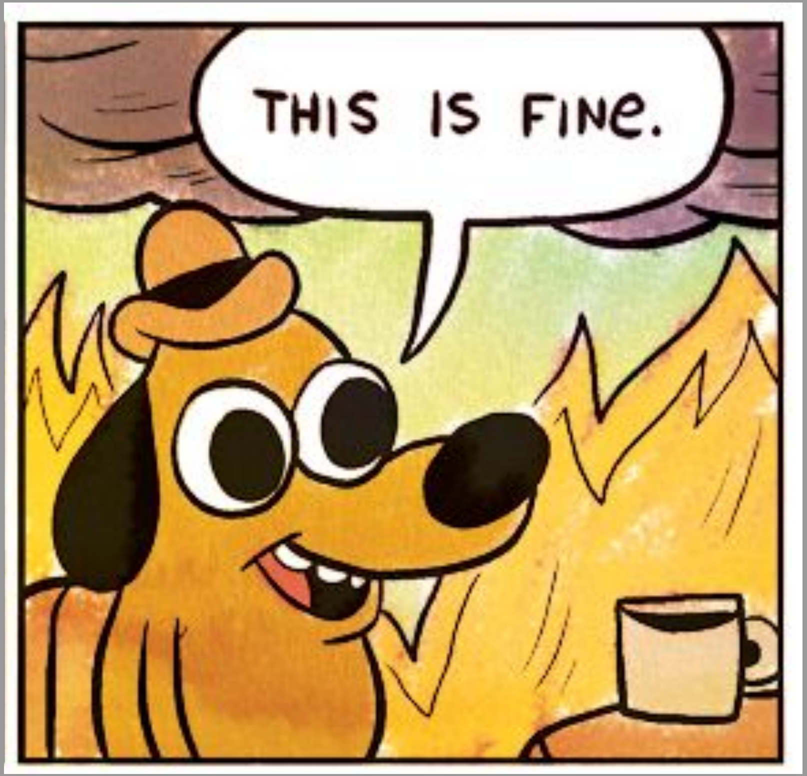


Total observations: 40,214

Support Vector Classifier

- Fit model overnight on <u>paid</u> AWS
- Kernel: ‘linear’
- Probability: True

- Best score approx .75
- Accidental termination without downloading, I have no proof.



Random Forest Classifier

- Fit model <= 25 mins
- Best score: .775
- Best params:
  - max_depth = 10
  - min_samples_leaf = 1
  - min_samples_split = 4
  - N_estimators = 150

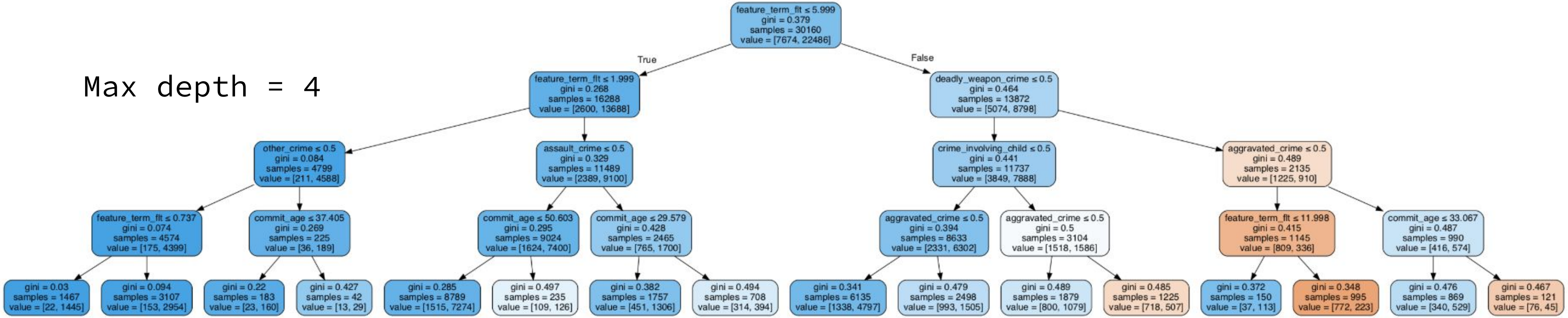
- Complete proof on my machine.

	Features
feature_term_flt	0.313261
commit_age	0.171880
deadly_weapon_crime	0.125341
aggravated_crime	0.104854
crime_involving_child	0.076614
assault_crime	0.045037
sex_crime	0.036206
theft_crime	0.035315
manslaughter_crime	0.032380
drugs_crime	0.028337
alcohol_crime	0.006767
vehicle_crime	0.006604
forgery_crime	0.005470

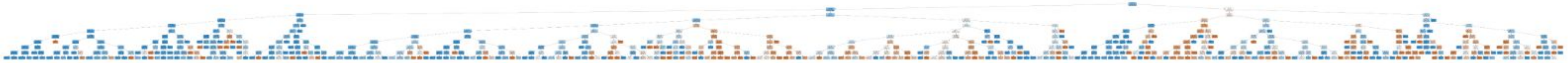
```
model.feature_importances_
```

Random Forest - One Decision Tree

Max depth = 4



Max depth = 10



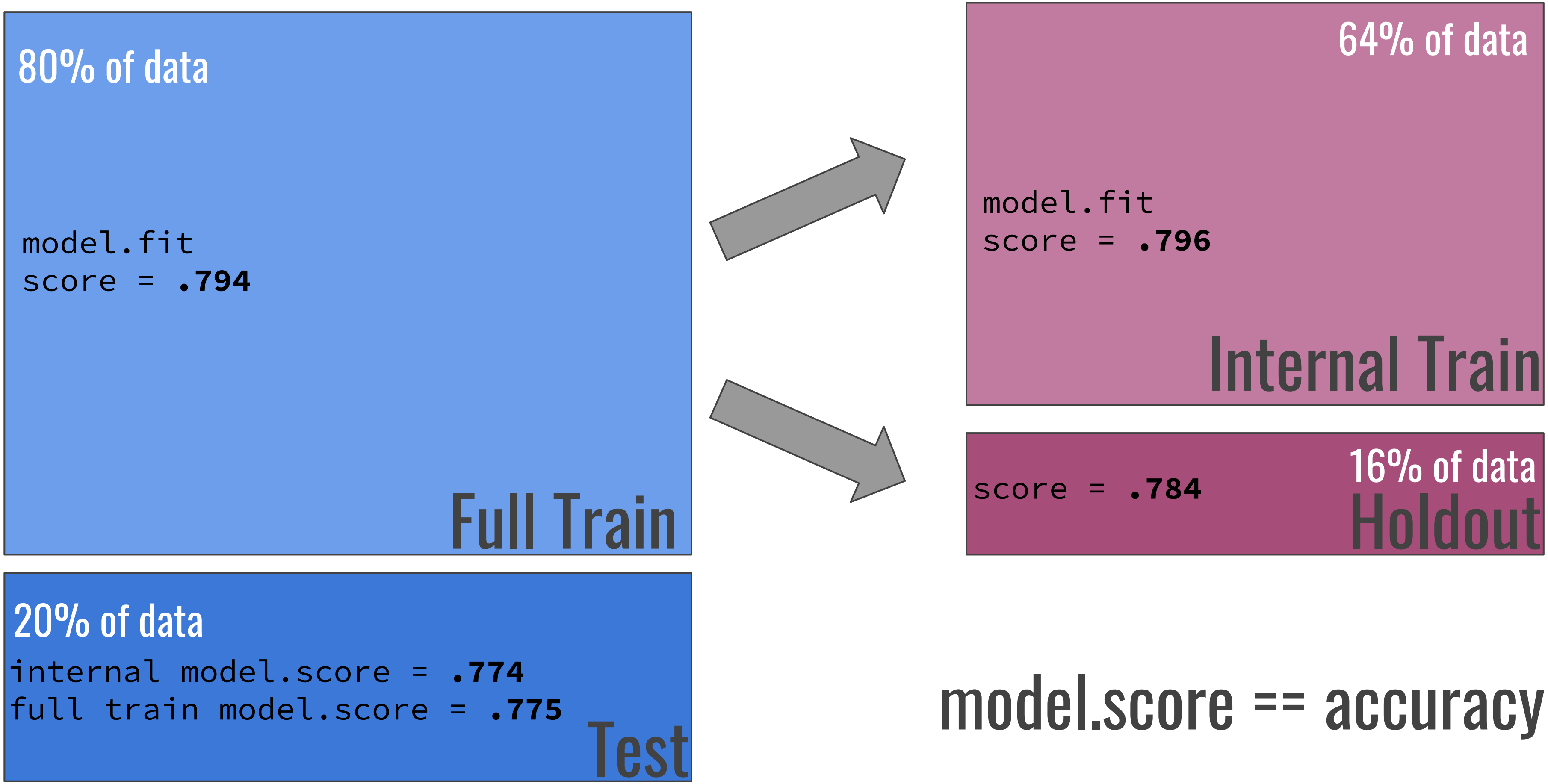
Logistic Regression

- Fit model <= 3 mins
- Best score: .768
- Best params:
  - C = 0.1

- Complete proof on my machine.

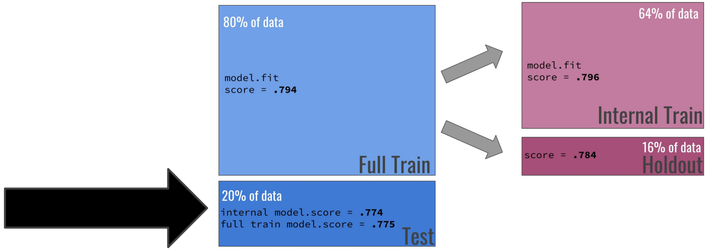
	Coefficients	Abs_Value
manslaughter_crime	-1.667695	1.667695
forgery_crime	1.575078	1.575078
credit_card_crime	1.072515	1.072515
deadly_weapon_crime	-0.798021	0.798021
vehicle_crime	0.759078	0.759078
drugs_crime	0.691760	0.691760
escape_crime	0.634955	0.634955
theft_crime	0.585717	0.585717
aggravated_crime	-0.571045	0.571045
tampering_crime	0.555448	0.555448
prison_related_crime	0.488335	0.488335
crime_involving_child	-0.429678	0.429678
elderly_crime	-0.304356	0.304356
arson_crime	-0.246226	0.246226
alcohol_crime	0.148257	0.148257
other_crime	0.124011	0.124011
sex_crime	-0.121948	0.121948
evading_crime	0.119084	0.119084
assault_crime	0.117495	0.117495
organized_crime	0.038085	0.038085

Random Forest Selected



model.score == accuracy

Random Forest - Metrics



	Predict Single Offenders	Predict Reoffenders
Actual Single Offenders	**650** TP: Single	**1908** FP: Multi
Actual Reoffenders	**354** FP: Single	**7142** TP: Multi

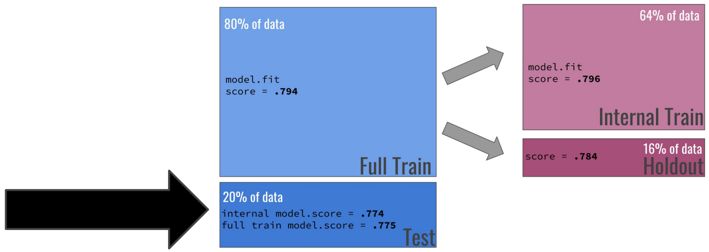
2,558

+ 7,496

Total: 10,054

*confusion matrix [inspo]()

Random Forest - Metrics



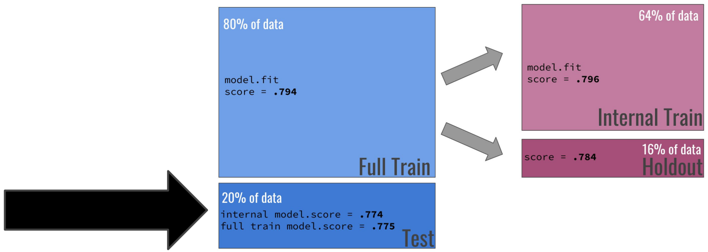
y classes

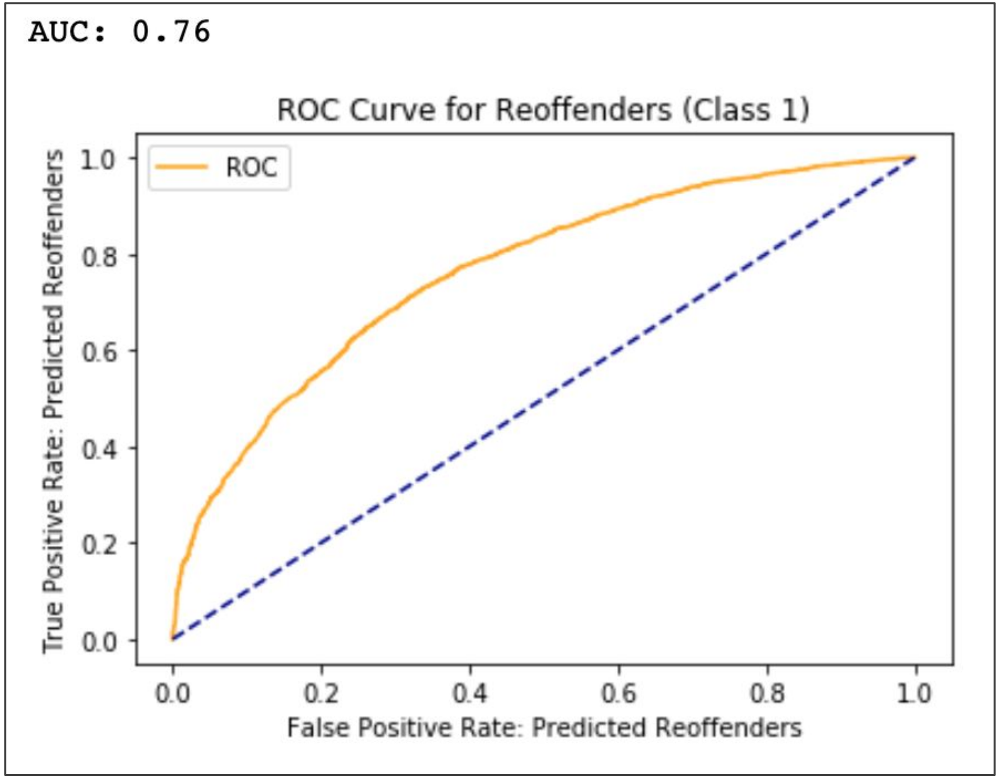
0	.254439
1	.745561

Accuracy: .775

	Precision	Recall
Class 0: Single Offenders	.65	.25
Class 1: Reoffenders	.79	.95

Random Forest - Metrics



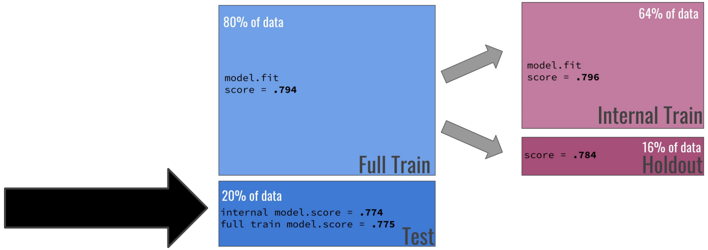


Test Data Set Statistics

Predictions and the people:

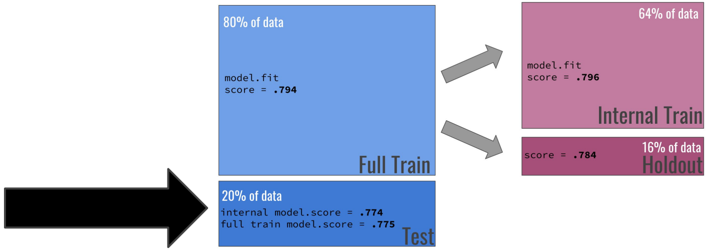
1. Looking at the numbers
2. Range of probability
3. Gender/Racial distributions
4. Noteworthy findings

Looking at the Numbers



	Minimum Value	Maximum Value
Probabilities	.12	.98
Age (at time crime occurred)	11.4	76.5
Projected Release Date	Aug 2019	Aug 2049

Looking at the Numbers



[redacted]

TDCJ Criminal History

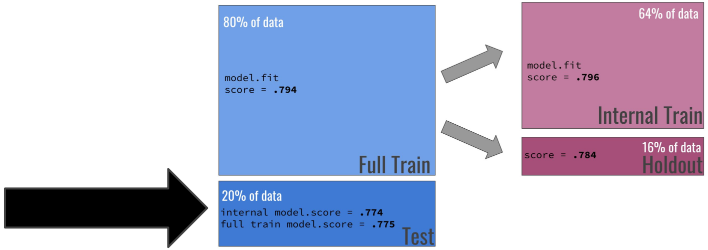
Crime	Committed On	County	Sentence	Sentence Began
FAILURE TO COMPLY W/REG REQ	3/22/2019	Bowie	1 year, 4 months	3/26/2019
HARASS PER CORR FAC	9/12/2008	Anderson	6 years	1/23/2013
INDEC W/CHILD	5/30/2002	Dallas	10 years	2/8/2004

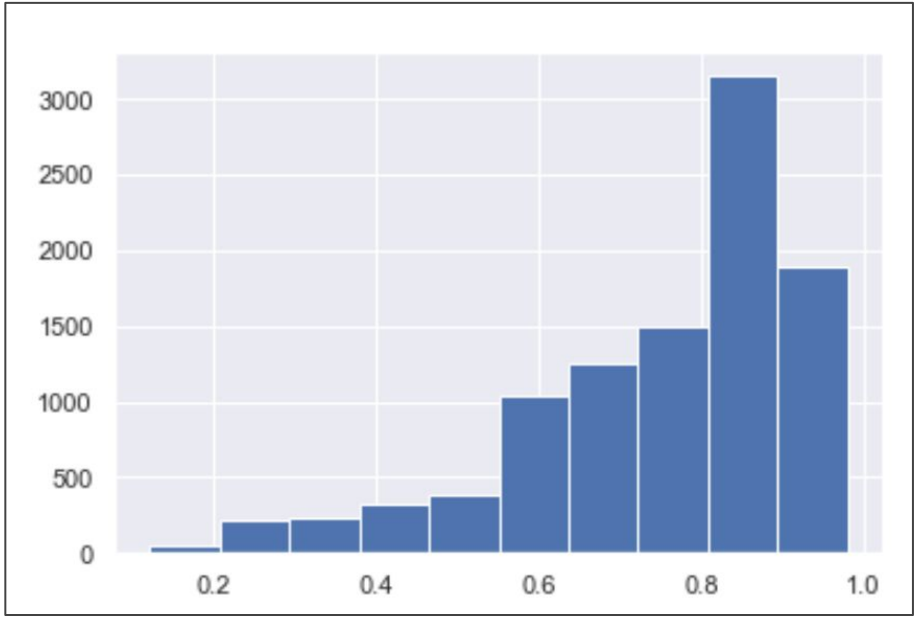
Maximum sentence	1 years, 4 months	**Sex**	Male
Projected release date	7/22/2020	**Age**	28
TDCJ ID	02265728	**Race**	White
Unit	Jester IV	**Height**	5 ft 8 in
DOB	1/1/1991	**Weight**	162 lbs
Home County	Bowie	**Hair Color**	Red
		Eye Color	Blue

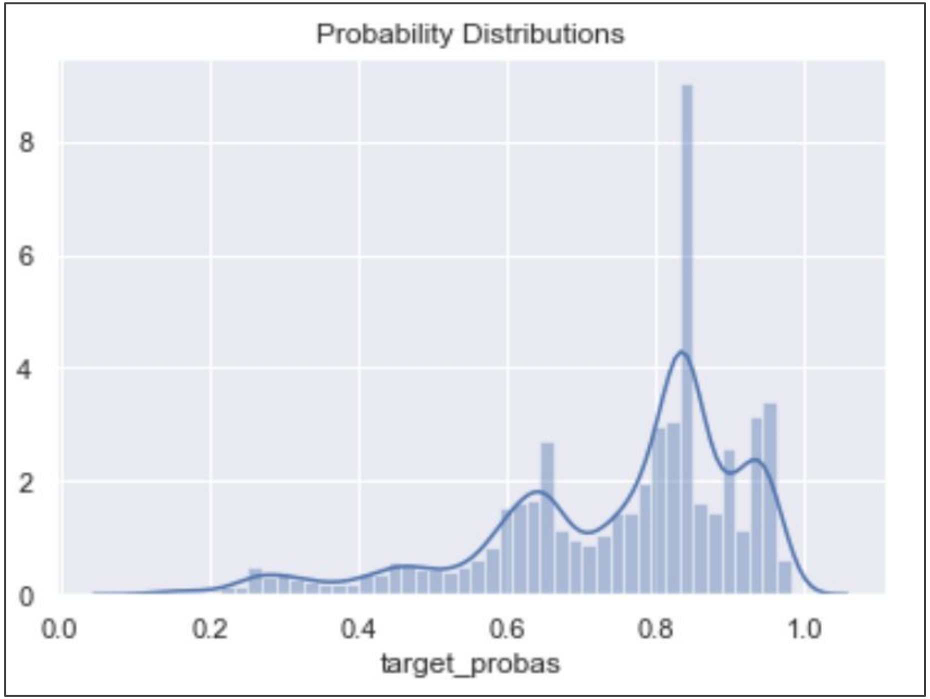
Age (at time crime occurred)

Minimum Value	Maximum Value
11.4	76.5

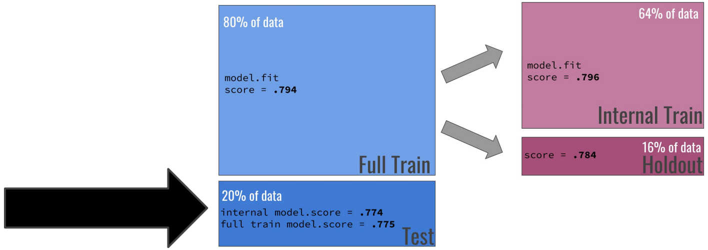
Probability Distributions





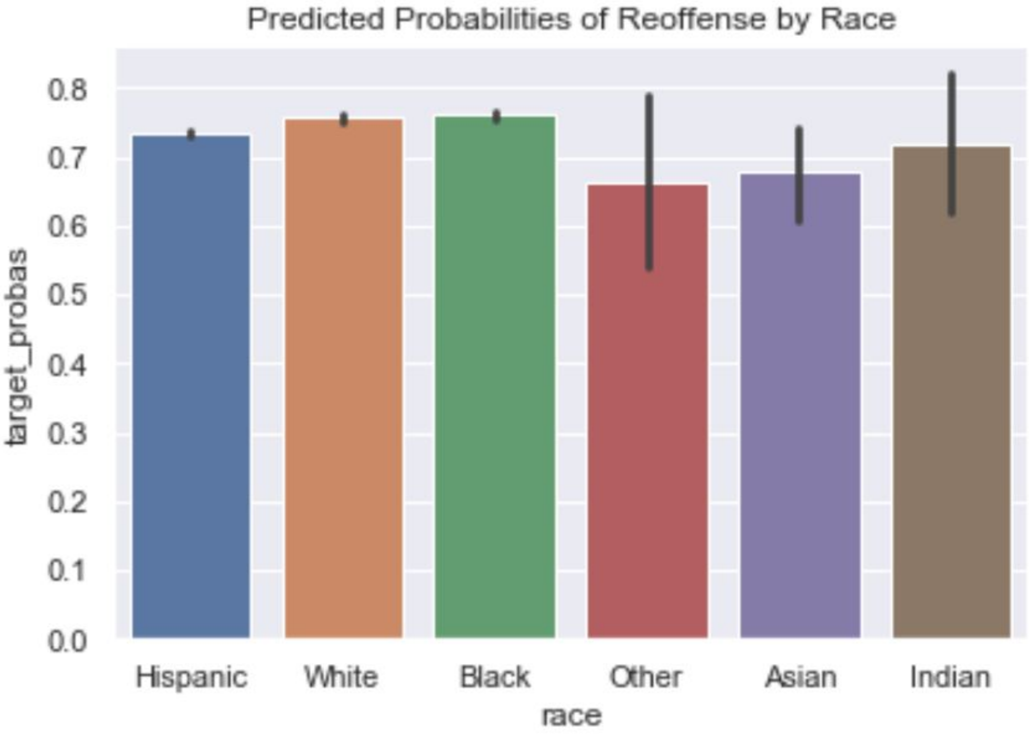


Probability Distributions



	race
Hispanic	3642
Black	3181
White	3179
Asian	39
Other	9
Indian	4

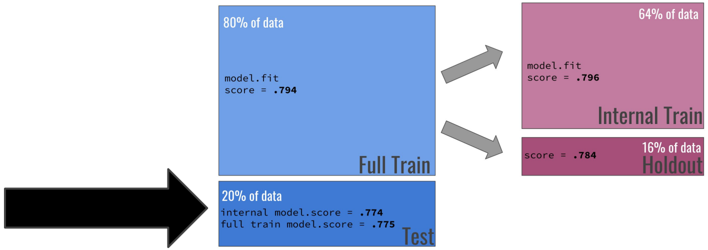
Counts

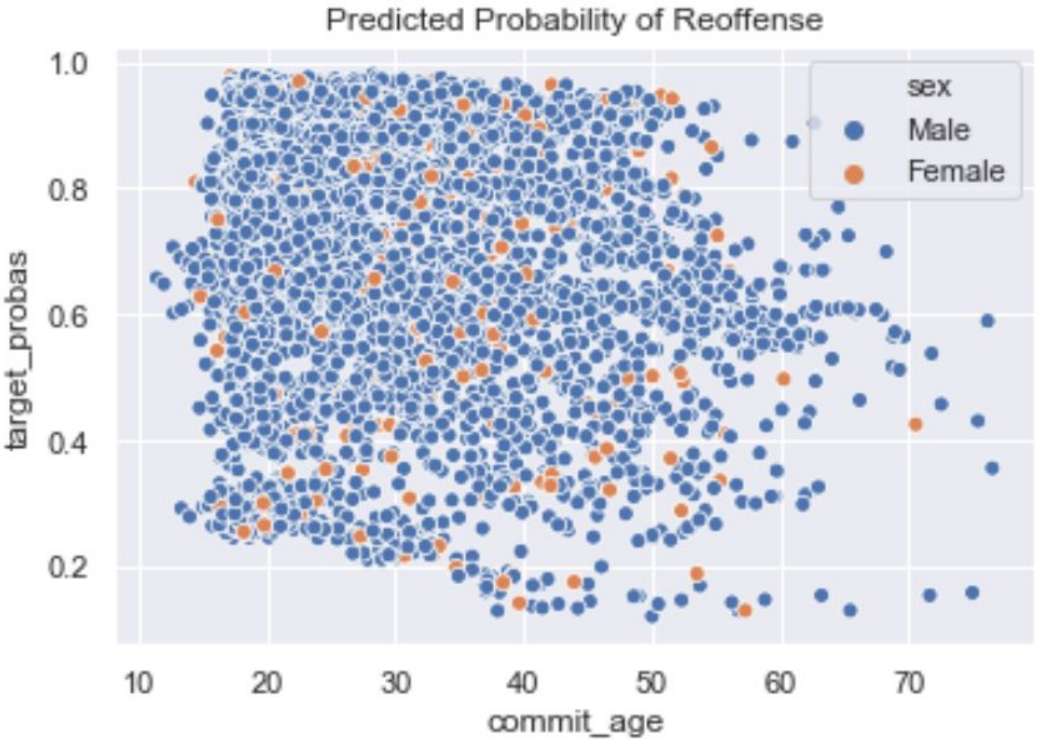


race	target_probas
Black	0.759784
White	0.755775
Hispanic	0.734207
Indian	0.718471
Asian	0.677388
Other	0.661585

Averages

Probability Distributions





	sex
Male	9352
Female	702

Counts

Conclusions

1. Practical Application
2. Additional considerations

Practical Application - Example with Test Dataset

1. Apply model to complete dataset of interest (ie by prison)

   Total observations: 10,054 inmates
   Probability Range: .12 - .98

2. Filter by probabilities of interest

   Total observations: 3003
   Probability Range: .30 - .70

3. Filter by projected release dates

   Total observations: 504
   Project Release Date Range: 2 years

Additional Considerations

- Incorporation of other features
- Learning more about options for feature engineering