Machine Learning to predict recidivism rates

By: David Troupe

Motivation

- United States has the world's largest prison population
- Recidivism is a major problem in the justice system
- Morals and Machine learning

Dataset

- National Corrections Reporting Program 1991 - 2014

File Dimensions:

- No. of Cases: 10,907,333
- No. of Variables: 18
- Record Length: 54
- Records per Case: 1
- Overall No. of Records: 10,907,333

Type of File: ASCII data file

Data Format: Logical record length

Variables within this Variable Group	
Variable	*Variable Label*
ABT_INMATE_ID	INMATE IDENTIFICATION ID
SEX	SEX OF INMATE
ADMTYPE	TYPE OF PRISON ADMISSION
OFFGENERAL	5-LEVEL CATEGORIZATION OF MOST SERIOUS SENTENCED OFFENSE
EDUCATION	HIGHEST LEVEL OF EDUCATION OF INMATE
ADMITYR	YEAR INMATE WAS ADMITTED TO PRISON
RELEASEYR	YEAR INMATE WAS RELEASED FROM PRISON
MAND_PRISREL_YEAR	YEAR OF MANDATORY PRISON RELEASE
PROJ_PRISREL_YEAR	YEAR OF PROJECTED PRISON RELEASE
PARELIG_YEAR	YEAR OF PAROLE ELIGIBILITY
SENTLGTH	MAXIMUM SENTENCE LENGTH FOR INMATE
OFFDETAIL	DETAILED CATEGORIZATION OF MOST SERIOUS SENTENCED OFFENSE
RACE	RACE/HISPANIC ETHNICITY OF INMATE
AGEADMIT	AGE AT ADMISSION
AGERELEASE	AGE AT RELEASE
TIMESRVD	TIME SERVED BY INMATE
RELTYPE	TYPE OF PRISON RELEASE
STATE	STATE WITH CUSTODY OF INMATE

Basic Analysis

Race:

35% African American

35% White

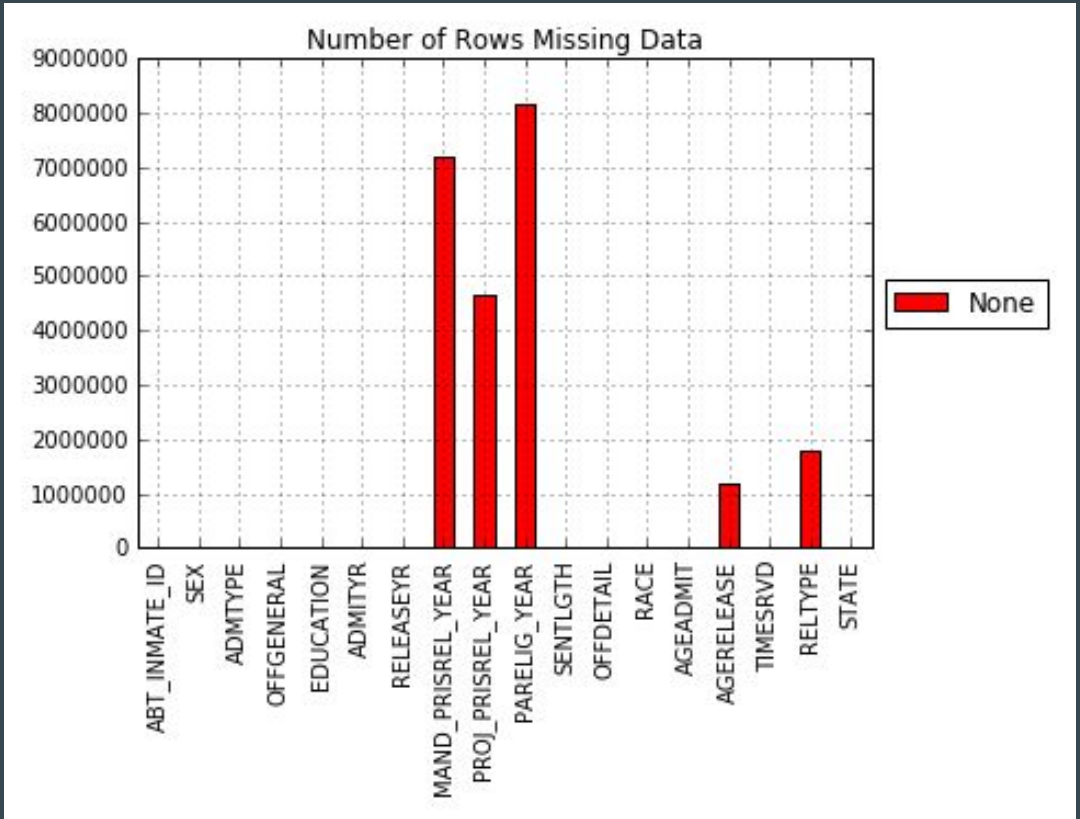
10% Missing

Repeat Offenders:

6,981,739 or 64%

90.875% are Male

Preprocessing Data - A lot of missing data



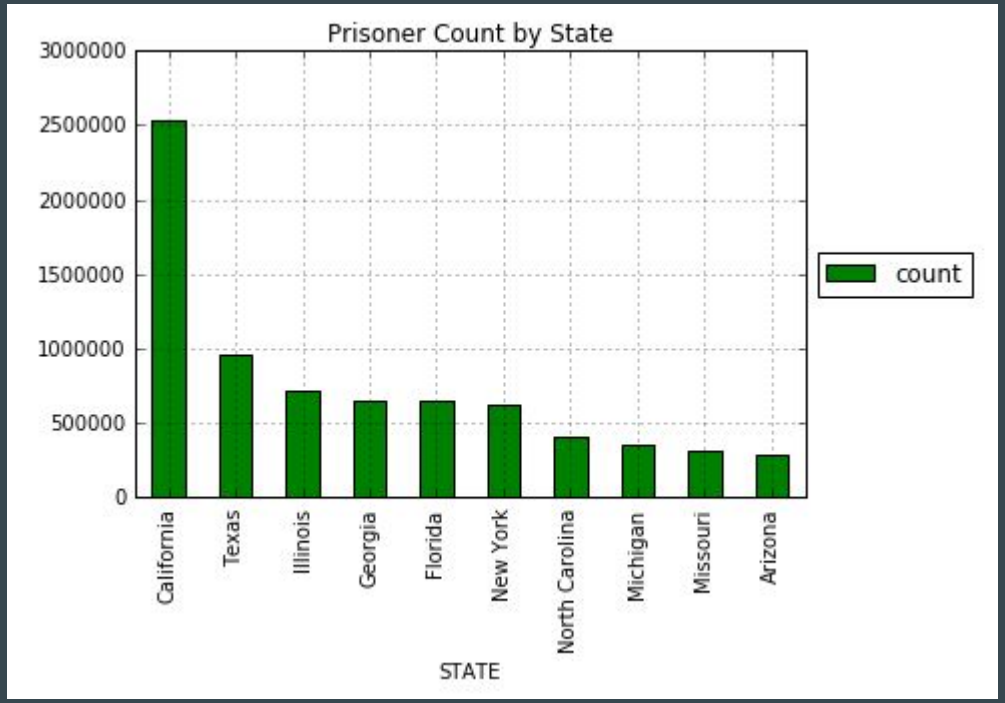
What to do?

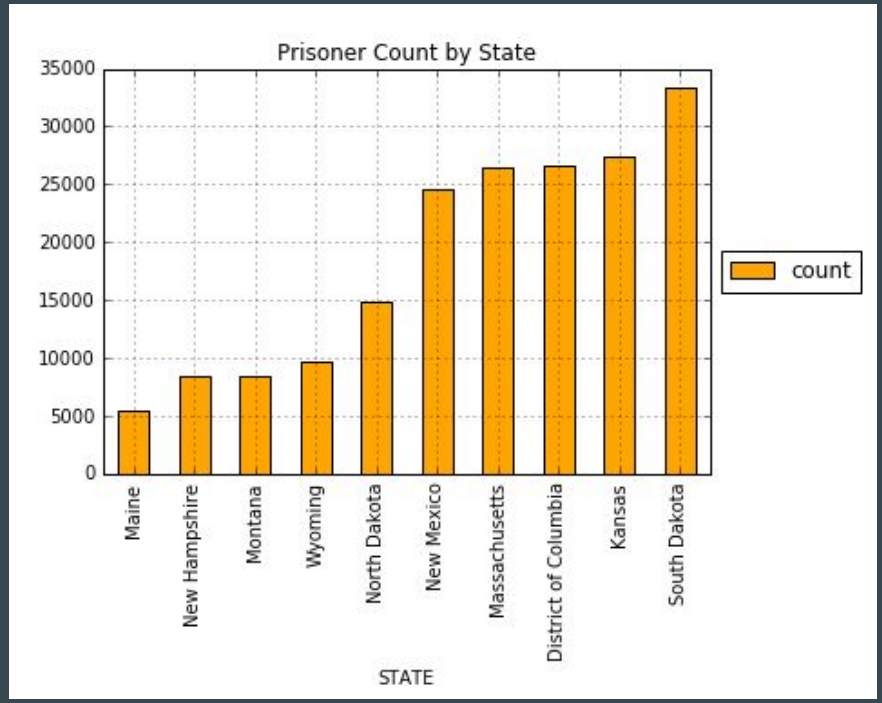
We can remove all the rows missing values by using df.dropna() but only 720,189 rows are complete or about 6% of the data is complete.

We will come back to this

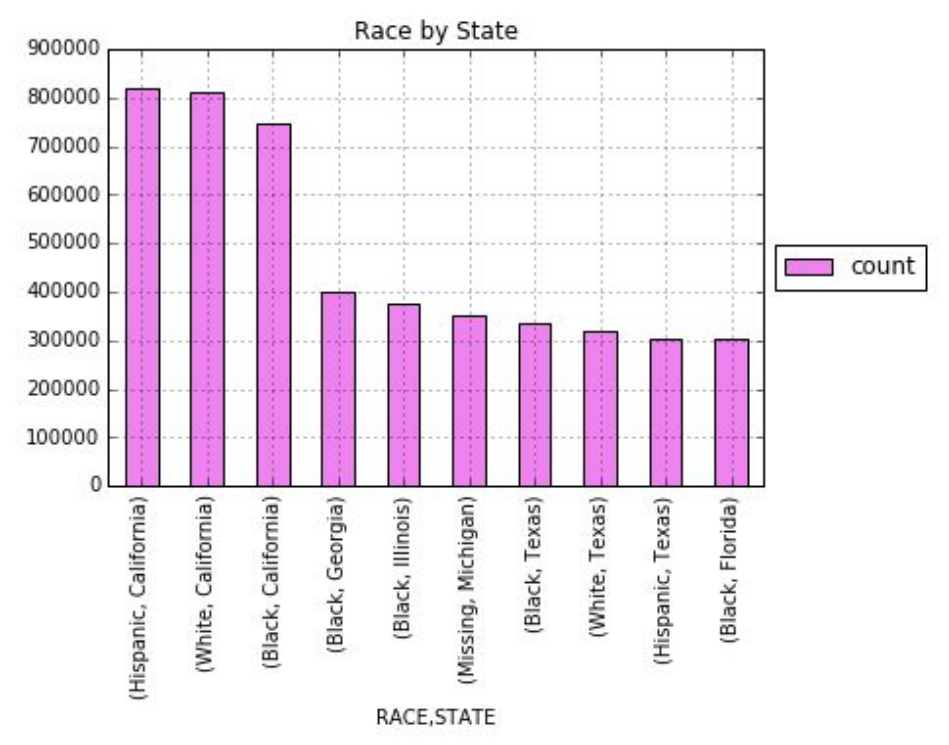
Imputing values is also an issue for some columns because more than half of the values are missing so it is unlikely that the imputed values with represent the data correctly

Prisoners by State



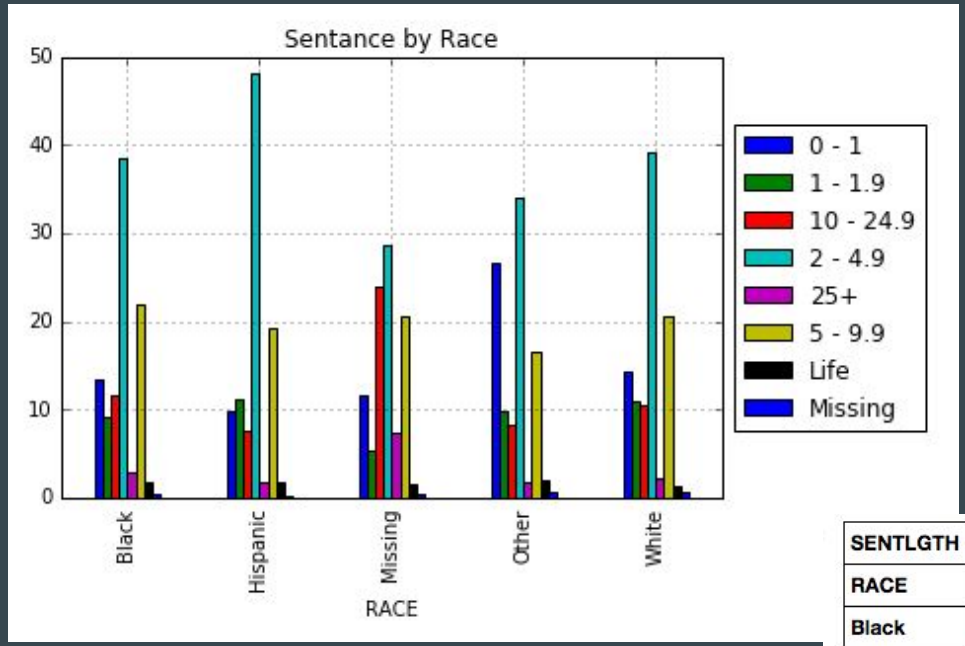


Race by State

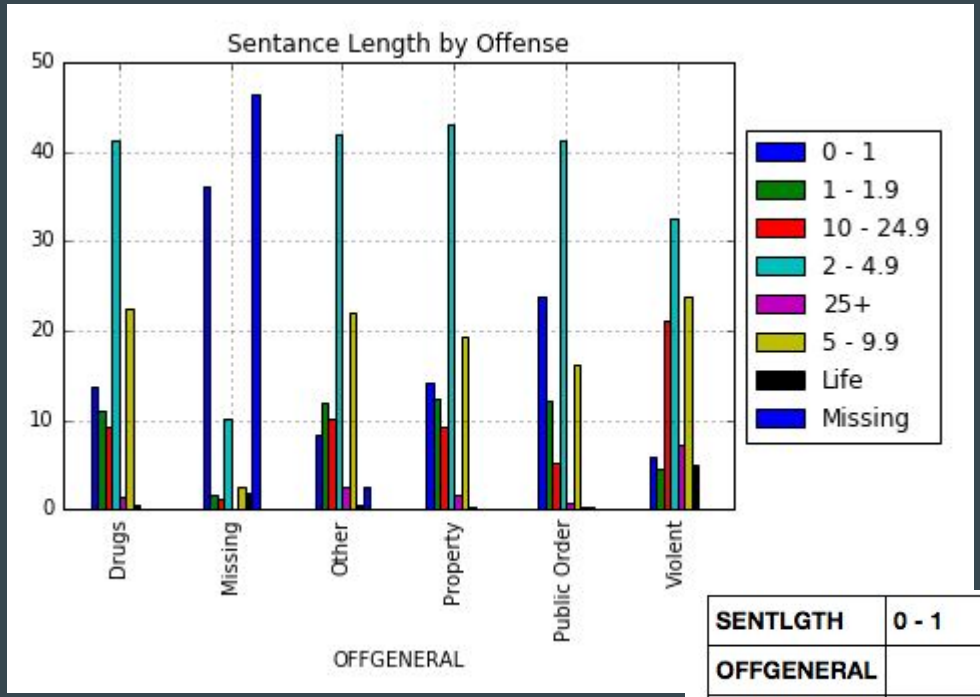


About the same number of hispanic, black, and white prisoners in the largest states. However, as percentages these groups are overrepresented.

Sentence Length by Race



SENTLGTH	0 - 1	1 - 1.9	10 - 24.9	2 - 4.9	25+	5 - 9.9	Life	Missing
RACE								
Black	13.452908	9.101673	11.580064	38.617899	2.968310	22.018191	1.752757	0.508198
Hispanic	9.946263	11.103155	7.653948	48.102016	1.878019	19.360450	1.775519	0.180630
Missing	11.614201	5.339084	24.076792	28.722387	7.404945	20.686627	1.685146	0.470818
Other	26.586445	9.963377	8.280232	33.987095	1.861118	16.542577	2.107061	0.672095
White	14.295786	10.993842	10.488890	39.322407	2.232418	20.728668	1.299821	0.638169

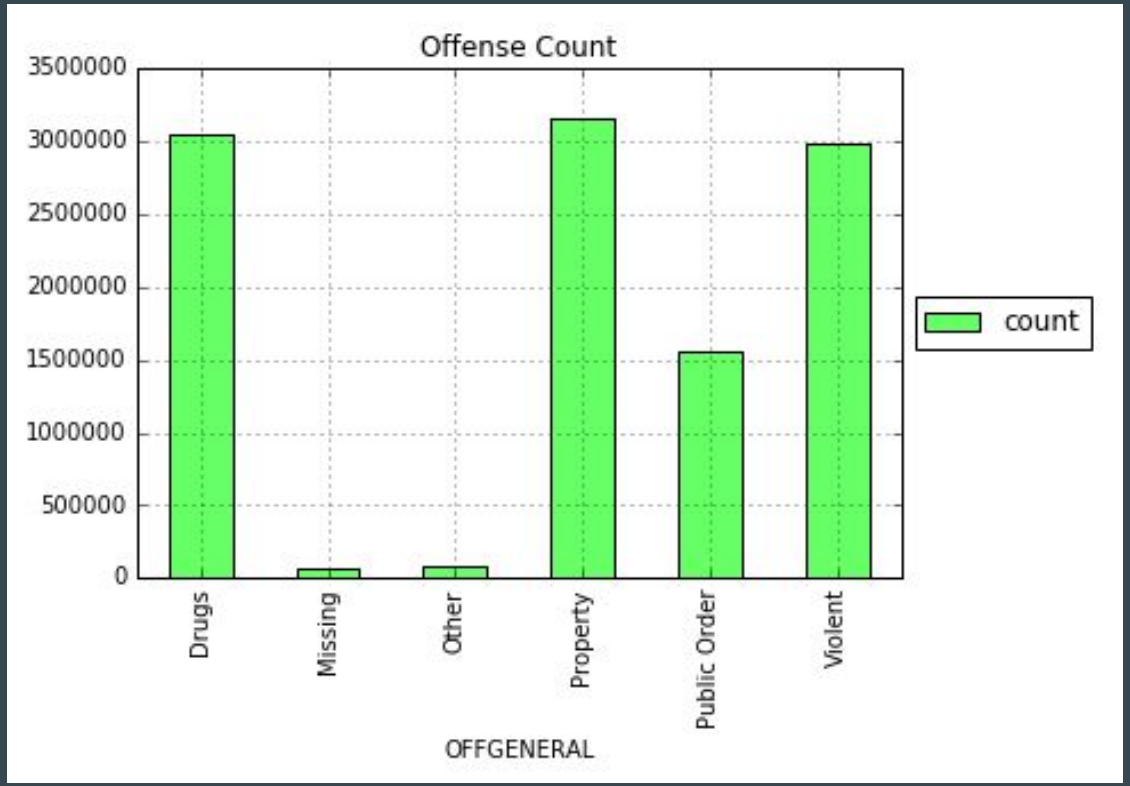


Sentence Length by Offense

Violent vs. Drugs

SENTLGTH	0 - 1	1 - 1.9	10 - 24.9	2 - 4.9	25+	5 - 9.9	Life	Missing
OFFGENERAL								
Drugs	13.681003	11.017551	9.276840	41.358339	1.513696	22.494780	0.520735	0.137056
Missing	36.080674	1.692910	1.192924	10.063097	0.170418	2.502746	1.856286	46.440945
Other	8.303531	11.916562	10.197440	41.900578	2.533261	22.089854	0.564608	2.494164
Property	14.268174	12.343258	9.165766	43.076718	1.575717	19.243617	0.195987	0.130764
Public Order	23.732960	12.088864	5.286476	41.321555	0.827589	16.183378	0.190258	0.368921
Violent	5.788639	4.459756	21.032803	32.595915	7.209000	23.785353	4.952593	0.175939

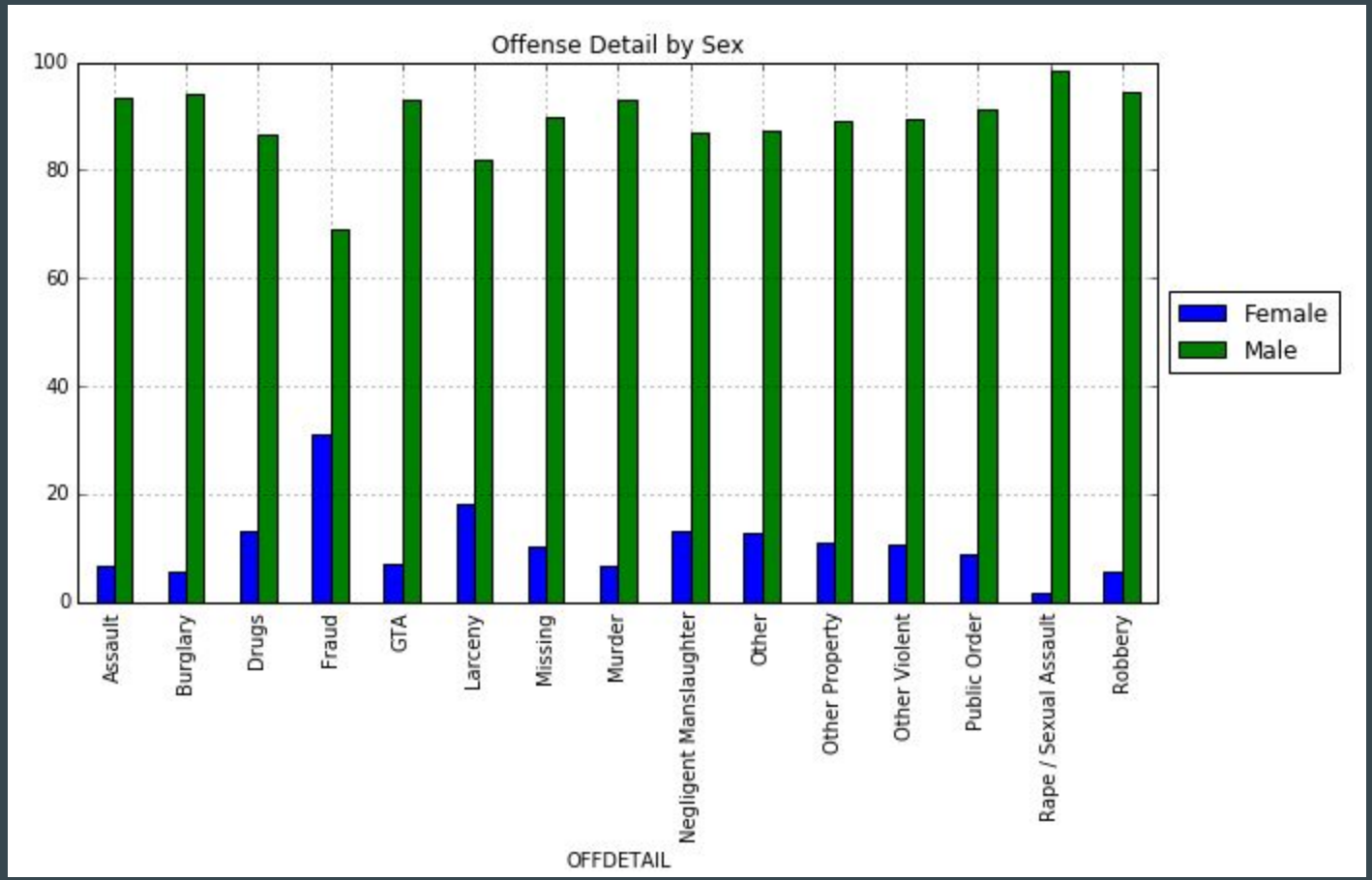
Violent vs Drugs



Violent Vs Drugs

SENTLGTH	0 - 1	1 - 1.9	10 - 24.9	2 - 4.9	25+	5 - 9.9	Life	Missing
OFFDETAIL								
Assault	10.826631	6.337671	11.659781	47.298930	2.328681	20.689201	0.742615	0.116491
Burglary	8.809195	8.301510	13.721045	41.560686	2.621558	24.496743	0.362520	0.126744
Drugs	13.681003	11.017551	9.276840	41.358339	1.513696	22.494780	0.520735	0.137056
Fraud	19.501422	13.416317	8.175684	40.064343	1.260291	17.347805	0.064613	0.169525
GTA	10.739043	17.535598	3.326040	51.872265	0.715427	15.708828	0.048274	0.054525
Larceny	20.193848	15.318514	5.302954	44.517715	0.695440	13.756000	0.082572	0.132956
Missing	36.080674	1.692910	1.192924	10.063097	0.170418	2.502746	1.856286	46.440945
Murder	0.878522	0.237286	26.744420	2.720663	22.228931	10.963867	35.414526	0.811785
Negligent Manslaughter	2.391671	2.950140	34.759673	26.265064	5.491870	27.425318	0.652836	0.063427
Other	8.303531	11.916562	10.197440	41.900578	2.533261	22.089854	0.564608	2.494164
Other Property	16.886083	14.570118	7.465466	41.991962	0.903251	17.892578	0.137839	0.152704
Other Violent	9.521945	10.280867	13.196009	38.718675	4.298360	21.392809	2.450989	0.140346
Public Order	23.732960	12.088864	5.286476	41.321555	0.827589	16.183378	0.190258	0.368921
Rape / Sexual Assault	2.600480	3.398252	30.529121	21.902889	10.873597	26.723712	3.811843	0.160106
Robbery	3.049892	3.001414	24.703500	31.475919	6.366986	29.899008	1.439824	0.063456

Men Vs Women



10.5 % Female

90.5 % Male

Repeat Offenders

64% are repeat offenders. Let's see if we can figure out what characteristics are the best predictors of whether someone will be a repeat offender.

How?

Logistic Regression and K Nearest Neighbors

Predicting - Repeat Offenders

1. Split data into training and test sets
2. Standardize the values (in this case not all numbers are in the same range)
3. Run logistical regression to calculate the accuracy

I did this on several samples of the data including random samples of varying size and noticed that if the data is too small the accuracy will be higher than it should

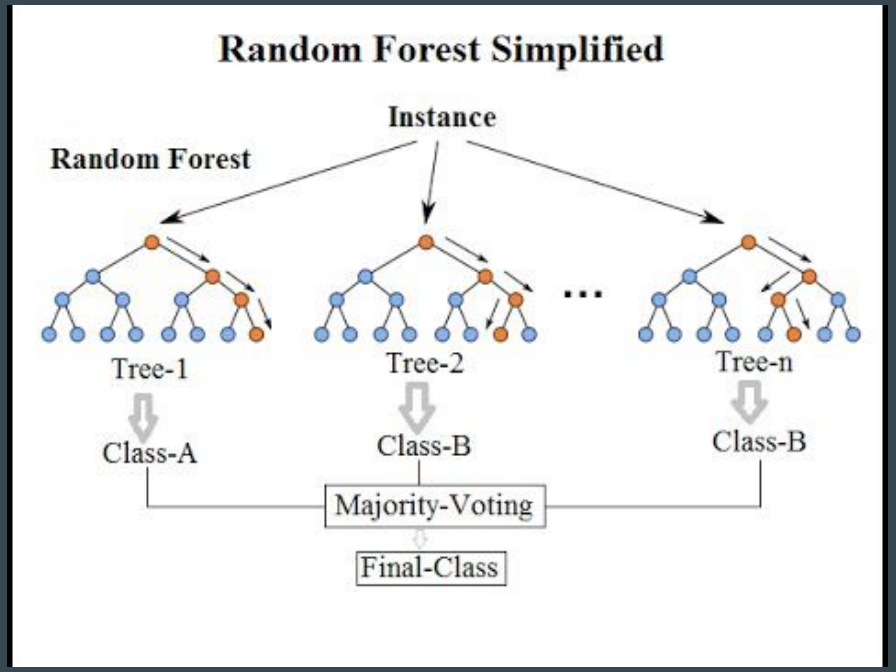
```
Training accuracy: 0.643916090969
Test accuracy: 0.643437043013
```

Disappointing Results

At this point I was disappointed with the results so I decided to explore the importance of each feature. The dataset currently has 16 dimensions.

To evaluate each variable I used a random forest to see which variables are the most important to predict repeat offenders.

Random Forest



from sklearn.ensemble import RandomForestClassifier

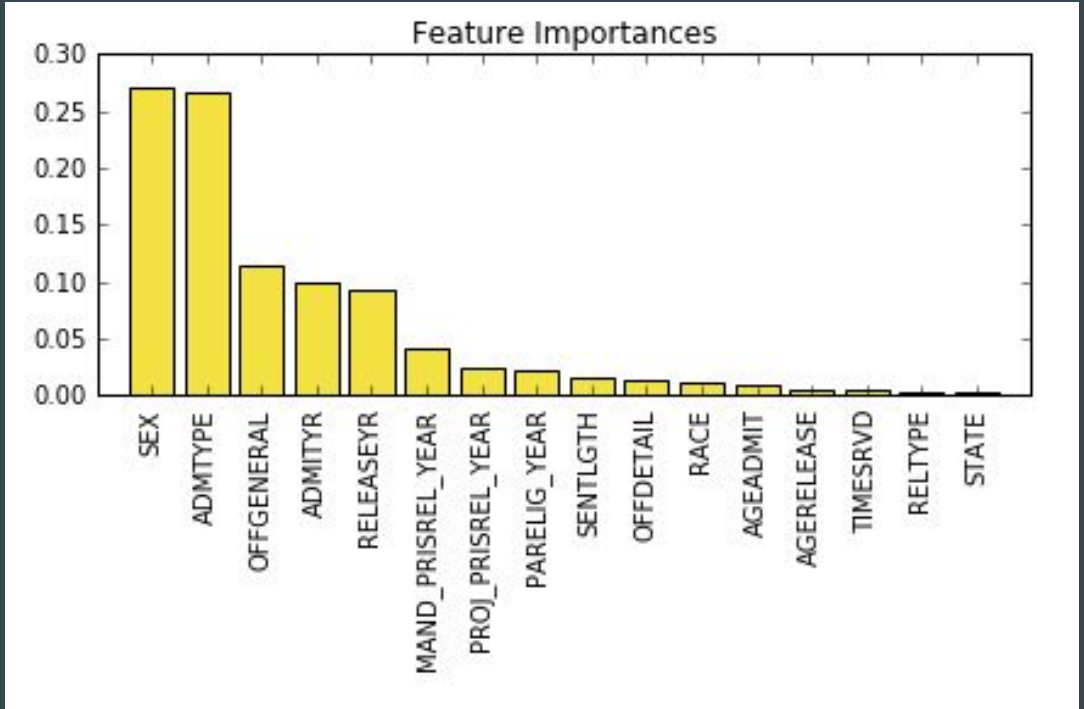
feat_labels = complete_rows.columns[1:]

forest = RandomForestClassifier(n_estimators=2000, max_depth=5, random_state=0, n_jobs=1)

Note: The number of estimators and the max_depth have a HUGE impact on how long this takes to run. Certainly not possible to run this on the entire dataset

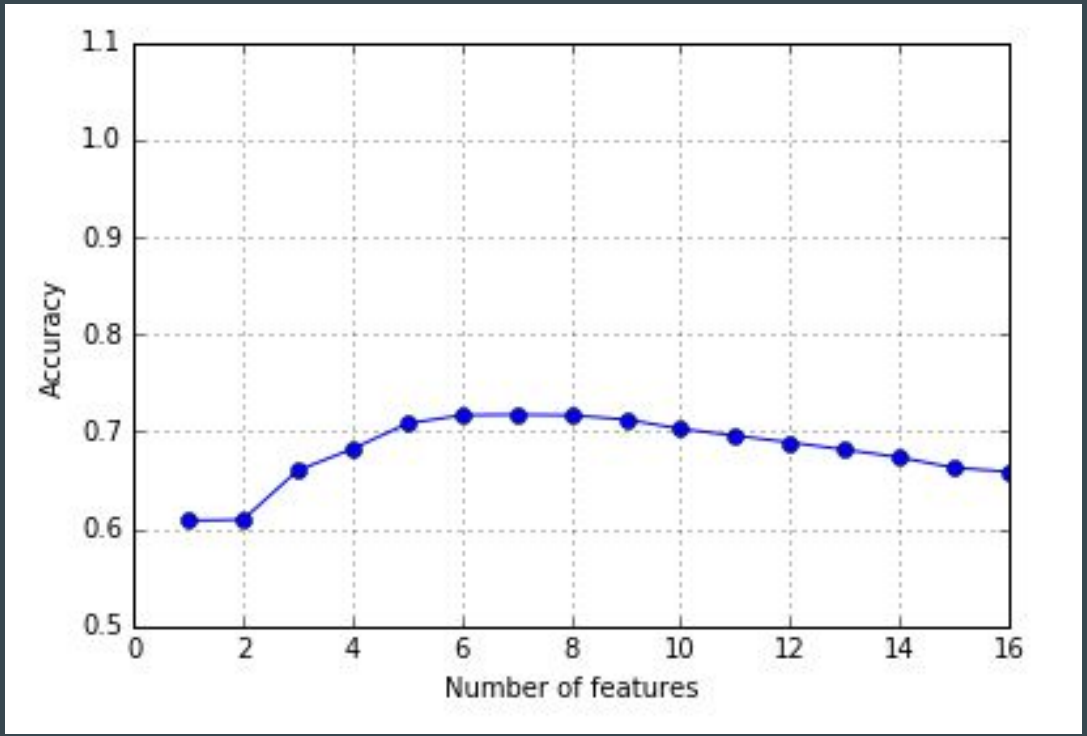
Feature Importance using Random Forest

1)	SEX	0.271421
2)	ADMTYPE	0.265530
3)	OFFGENERAL	0.114838
4)	ADMITYR	0.100254
5)	RELEASEYR	0.092680
6)	MAND_PRISREL_YEAR	0.042345
7)	PROJ_PRISREL_YEAR	0.024354
8)	PARELIG_YEAR	0.021873
9)	SENTLGTH	0.015724
10)	OFFDETAIL	0.014365
11)	RACE	0.010992
12)	AGEADMIT	0.008765
13)	AGERELEASE	0.005322
14)	TIMESRVD	0.004487
15)	RELTYPE	0.004090
16)	STATE	0.002960



Clearly we can greatly reduce the dimensionality of this set

Feature Importance using KNN & Sequential Backward Selection



With k = 2

6 Features For Best Results:

1. ADMTYPE
2. RELEASEYR
3. MAND_PRISREL_YEAR
4. PARELIG_YEAR
5. TIMESRVD
6. STATE

Thus we can remove 10 dimensions! This provides a slight improvement over logistic regression

Accuracy Results of Reducing Dimensions

Logistic Regression - Reduced to 6 dimensions and accuracy is only reduced by ~

```
Training accuracy: 0.639990923187
Test accuracy: 0.645942061871
```

vs

```
Training accuracy: 0.637769322991
Test accuracy: 0.636569924437
```

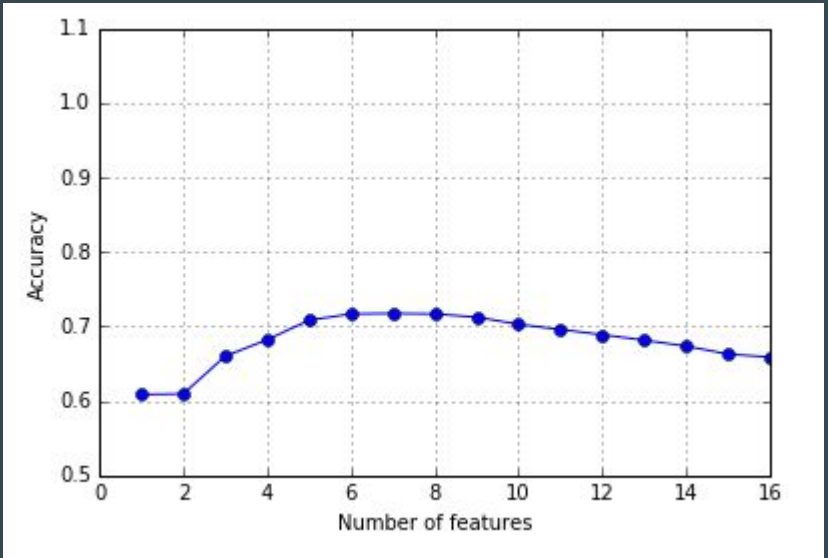
KNN - Reduced to 6 dimensions, increased n_neighbors to 7, and doubled the sample size

```
0.71741925724226607
```

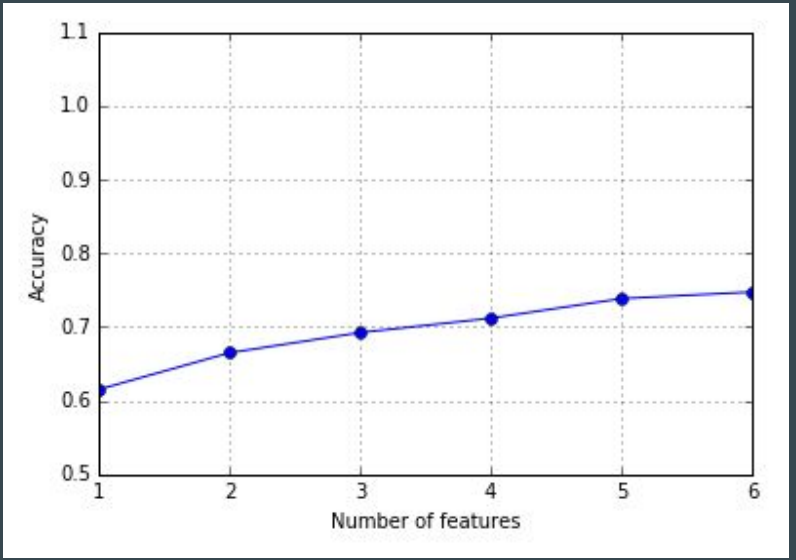
vs

```
0.74733378715679599
```

Performance Results of Reducing Dimensions



About 8 Minutes



~80 Seconds Running time

Conclusion

State Parole Boards Use Software to Decide Which Inmates to Release

Programs look at prisoners' biographies for patterns that predict future crime

Overall - My Algorithm has **74.7%** accuracy which is lower than I hoped for. I think with additional time, data, or the use of other algorithms I could make improvements.

Moral Dilemmas - Would it be acceptable to use this algorithm? When? How?

Already Happening - reported by WSJ

At least 15 States are currently using software known as "COMPAS"