US Crime Predictions

MSBA 273 - Group #9
Chengwu Weng
Mira Daya
Misha Khan
Yue Fang

Overview

Objective

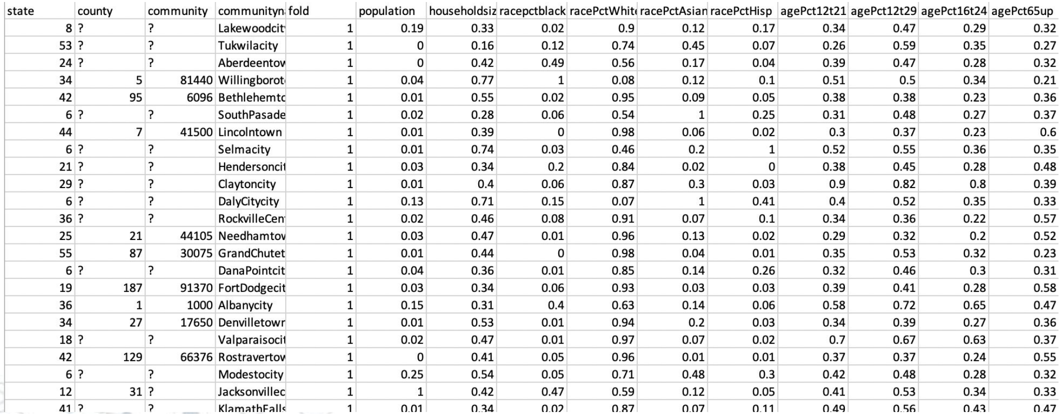
With our project, we examined the data collected throughout the nation and evaluated which certain attributes contribute to higher rates of violent crime.

Data Collection

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems



state	county	community	communityn	fold	population	householdsiz	racepctblack	racePctWhit	racePctAsian	racePctHisp	agePct12t21	agePct12t29	agePct16t24	agePct65up
8	?	?	Lakewoodcit	1	0.19	0.33	0.02	0.9	0.12	0.17	0.34	0.47	0.29	0.32
53	?	?	Tukwilacity	1	0	0.16	0.12	0.74	0.45	0.07	0.26	0.59	0.35	0.27
24	?	?	Aberdeentov	1	0	0.42	0.49	0.56	0.17	0.04	0.39	0.47	0.28	0.32
34	5	81440	Willingborot	1	0.04	0.77	1	0.08	0.12	0.1	0.51	0.5	0.34	0.21
42	95	6096	Bethlehemtc	1	0.01	0.55	0.02	0.95	0.09	0.05	0.38	0.38	0.23	0.36
6	?	?	SouthPasade	1	0.02	0.28	0.06	0.54	1	0.25	0.31	0.48	0.27	0.37
44	7	41500	Lincolntown	1	0.01	0.39	0	0.98	0.06	0.02	0.3	0.37	0.23	0.6
6	?	?	Selmacity	1	0.01	0.74	0.03	0.46	0.2	1	0.52	0.55	0.36	0.35
21	?	?	Hendersoncit	1	0.03	0.34	0.2	0.84	0.02	0	0.38	0.45	0.28	0.48
29	?	?	Claytoncity	1	0.01	0.4	0.06	0.87	0.3	0.03	0.9	0.82	0.8	0.39
6	?	?	DalyCitycity	1	0.13	0.71	0.15	0.07	1	0.41	0.4	0.52	0.35	0.33
36	?	?	RockvilleCen	1	0.02	0.46	0.08	0.91	0.07	0.1	0.34	0.36	0.22	0.57
25	21	44105	Needhamtov	1	0.03	0.47	0.01	0.96	0.13	0.02	0.29	0.32	0.2	0.52
55	87	30075	GrandChutet	1	0.01	0.44	0	0.98	0.04	0.01	0.35	0.53	0.32	0.23
6	?	?	DanaPointcit	1	0.04	0.36	0.01	0.85	0.14	0.26	0.32	0.46	0.3	0.31
19	187	91370	FortDodgecit	1	0.03	0.34	0.06	0.93	0.03	0.03	0.39	0.41	0.28	0.58
36	1	1000	Albanycity	1	0.15	0.31	0.4	0.63	0.14	0.06	0.58	0.72	0.65	0.47
34	27	17650	Denvilletowr	1	0.01	0.53	0.01	0.94	0.2	0.03	0.34	0.39	0.27	0.36
18	?	?	Valparaisocit	1	0.02	0.47	0.01	0.97	0.07	0.02	0.7	0.67	0.63	0.37
42	129	66376	Rostravertov	1	0	0.41	0.05	0.96	0.01	0.01	0.37	0.37	0.24	0.55
6	?	?	Modestocity	1	0.25	0.54	0.05	0.71	0.48	0.3	0.42	0.48	0.28	0.32
12	31	?	Jacksonvillec	1	1	0.42	0.47	0.59	0.12	0.05	0.41	0.53	0.34	0.33
41	?	?	KlamathFalls	1	0.01	0.34	0.02	0.87	0.07	0.11	0.49	0.56	0.43	0.47

- **What?**
  - ○ US Communities & Crime
  - ○ 128 attributes
- **When?**
  - ○ 1990 US Census
  - ○ 1990 US Law Enforcement
  - ○ 1995 FBI crime data

Data Overview

Data	Variable
Non-predictive Attributes	state
	county
	community
	communityname
	fold
Predictive Attributes	demographic (70)
	income (20)
	education (3)
	employment (6)
	police-related (21)
	crime-related (2)
Goal Attributes	“ViolentCrimesPerPop”

Data Cleaning

Step 1: Convert ?'s to NaN

If NaN values were larger than 50% of the column, we removed the variable from our dataset.

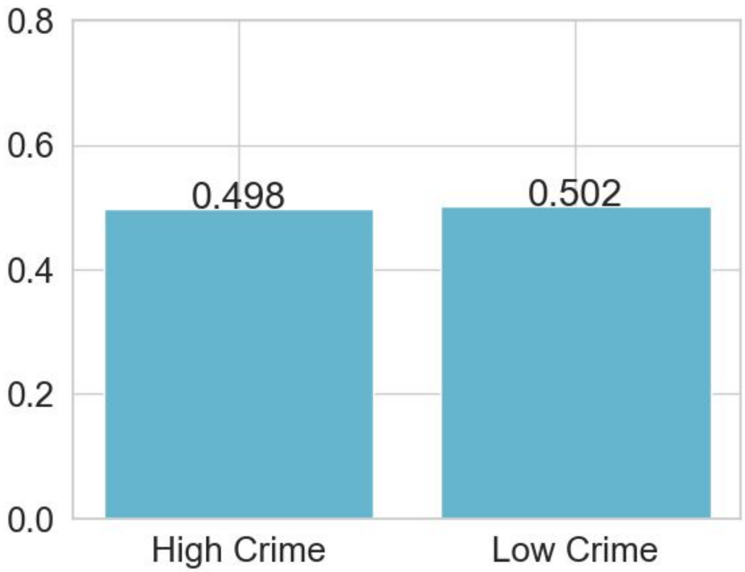
We eliminated 22 columns.

LemasSwFTPerPop	*PctPolicHisp*
LemasSwFTFieldOps	*PctPolicAsian*
LemasSwFTFieldPerPop	*PctPolicMinor*
LemasTotalReq	*OfficAssgnDrugUnits*
LemasTotReqPerPop	*NumKindsDrugsSeiz*
PolicReqPerOffic	*PolicAveOTWorked*
PolicPerPop	*PolicCars*
RacialMatchCommPol	*PolicOperBudg*
PctPolicWhite	*LemasPctPolicOnPatr*
PctPolicBlack	*LemasGangUnitDeploy*
LemasSwFTPerPop	*PolicBudgPerPop*

Preprocessing

Set threshold using median 0.15 to determine HighCrime rate

- HighCrime areas = 49.8%
- LowCrime areas = 50.2%

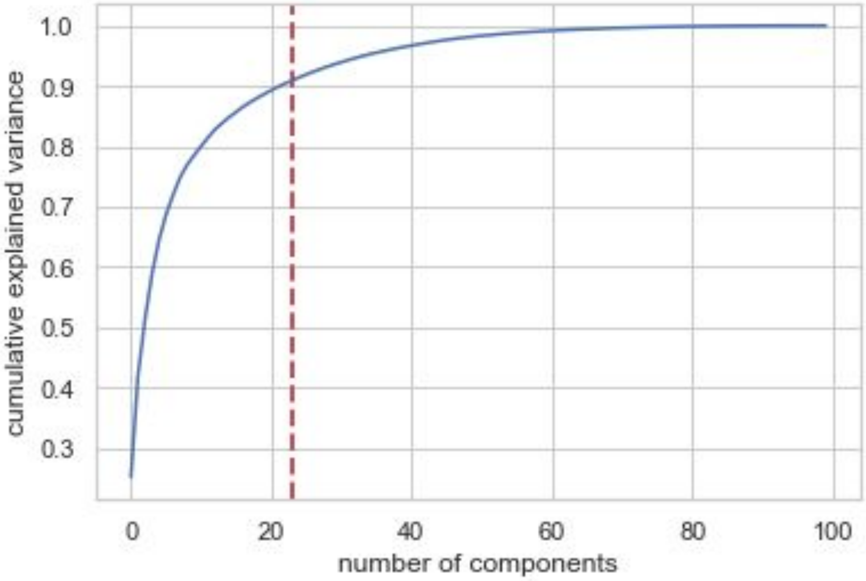


$$HighCrime = \begin{cases} 1, & if\ ViolentCrimesPrerPop\ \geq 0.15 \\ 0, & Otherwise \end{cases}$$

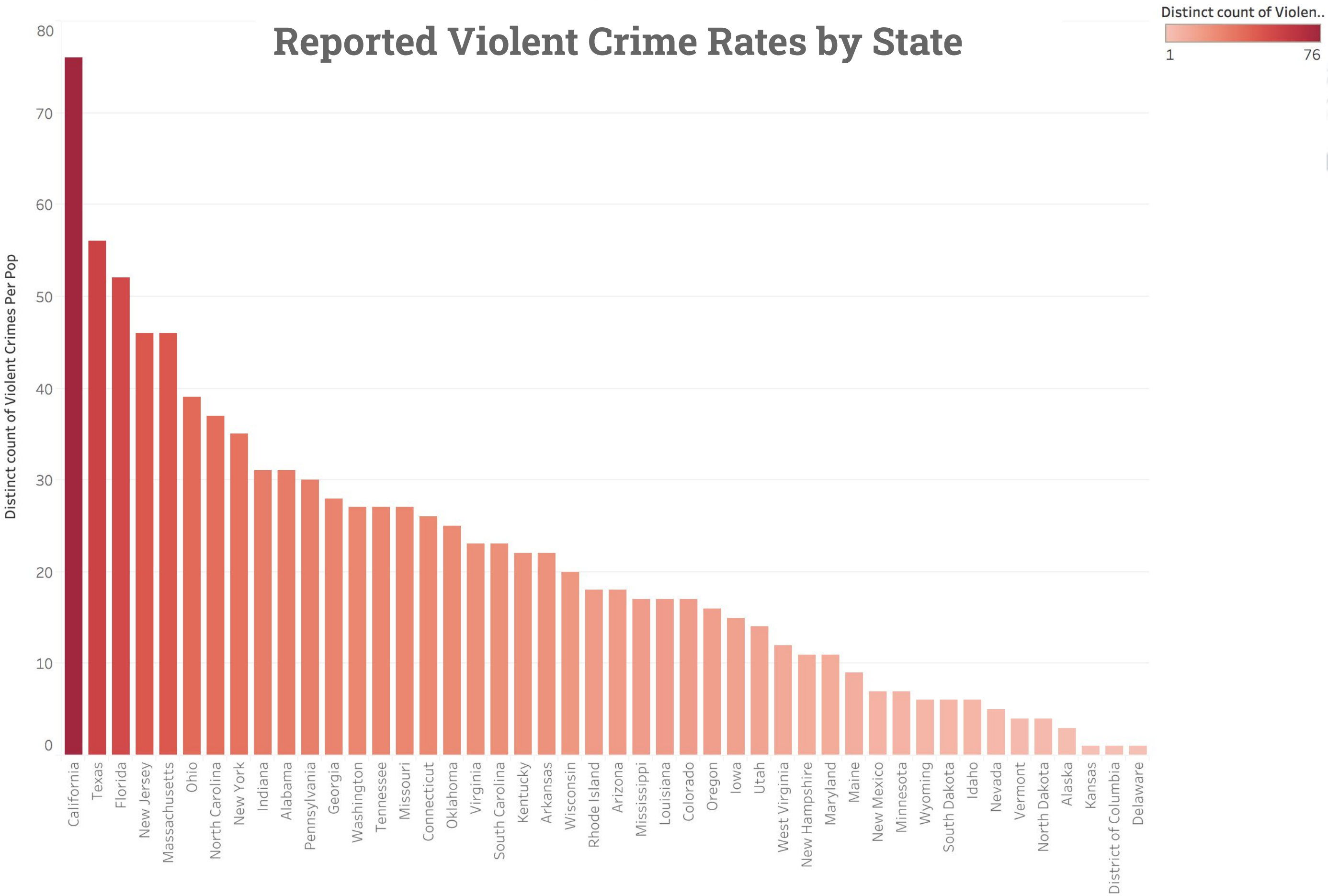
Preprocessing

Principal Component Analysis (PCA)

PCA is useful to reduce the number of variables from a large dataset and transforms a set of variables into a new set of uncorrelated variables.

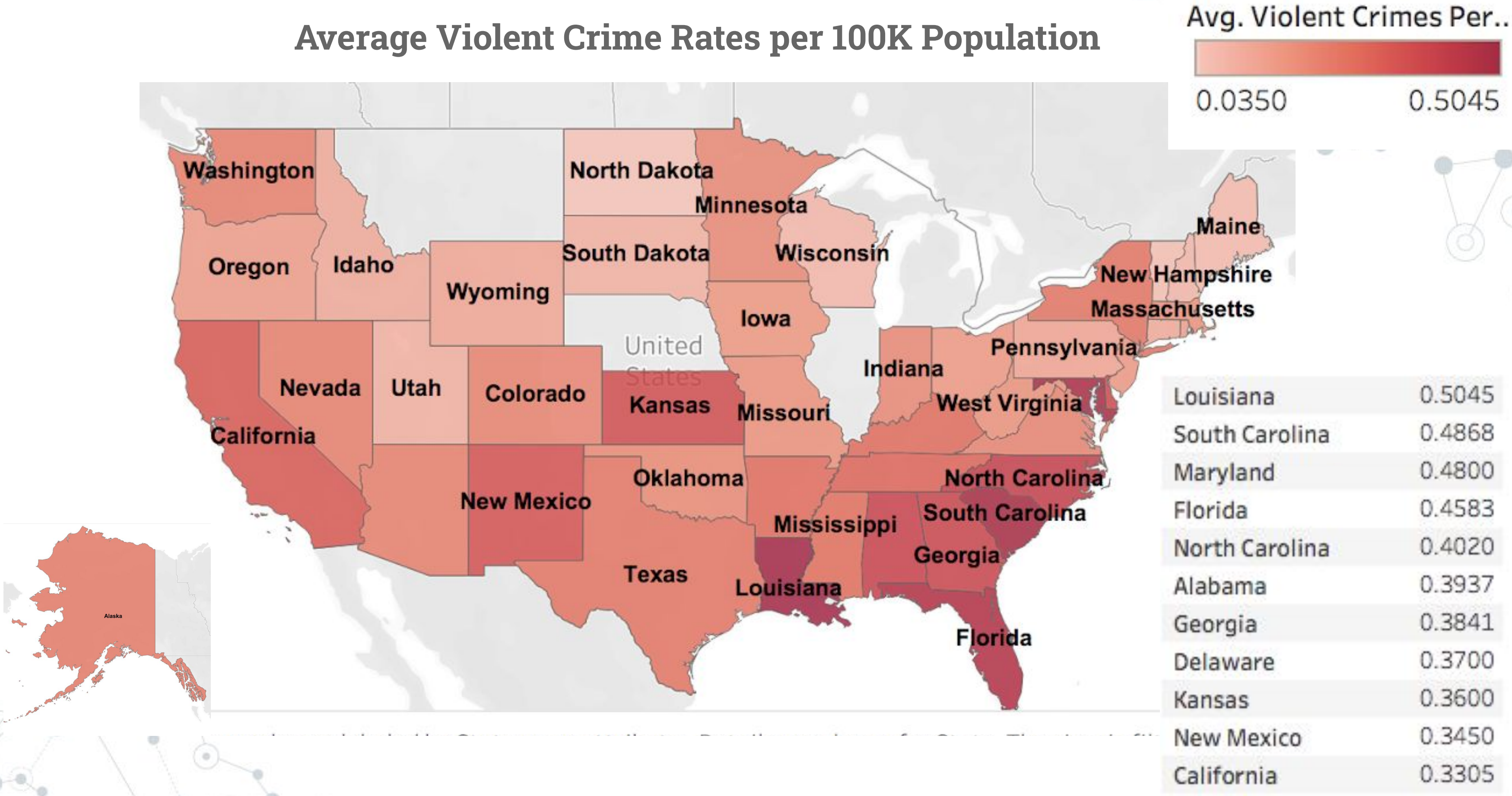


Data Visualization


Reported Violent Crime Rates by State

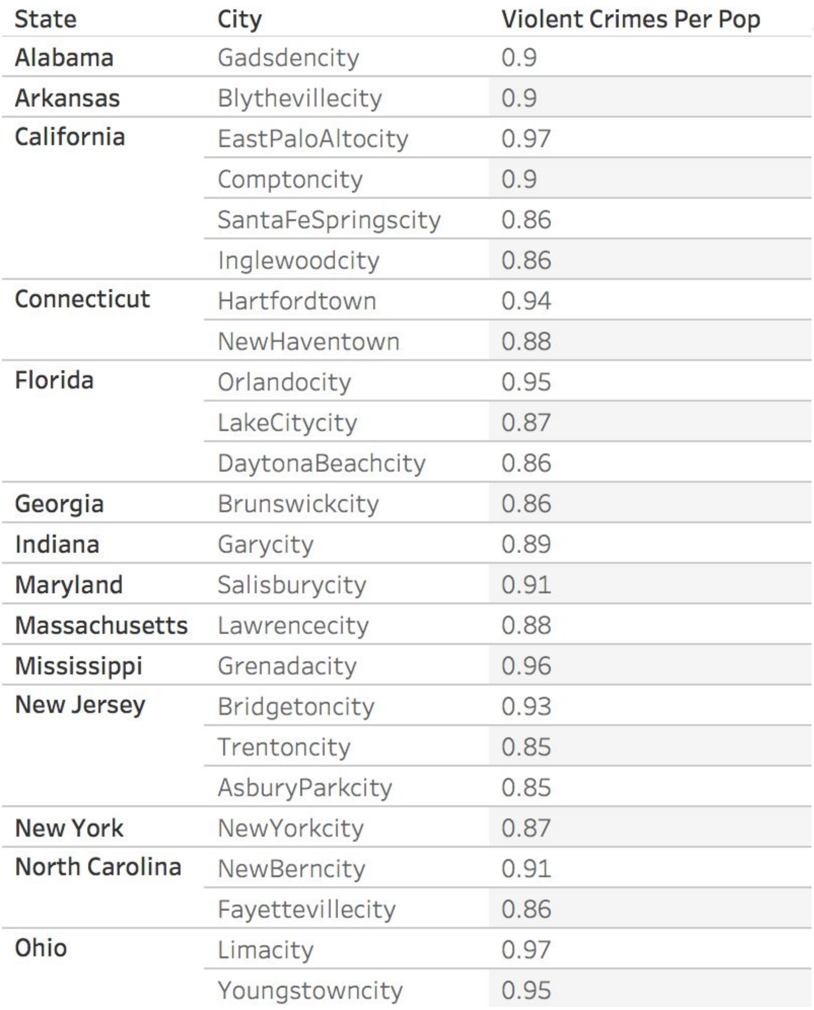
Data Visualization

Average Violent Crime Rates per 100K Population



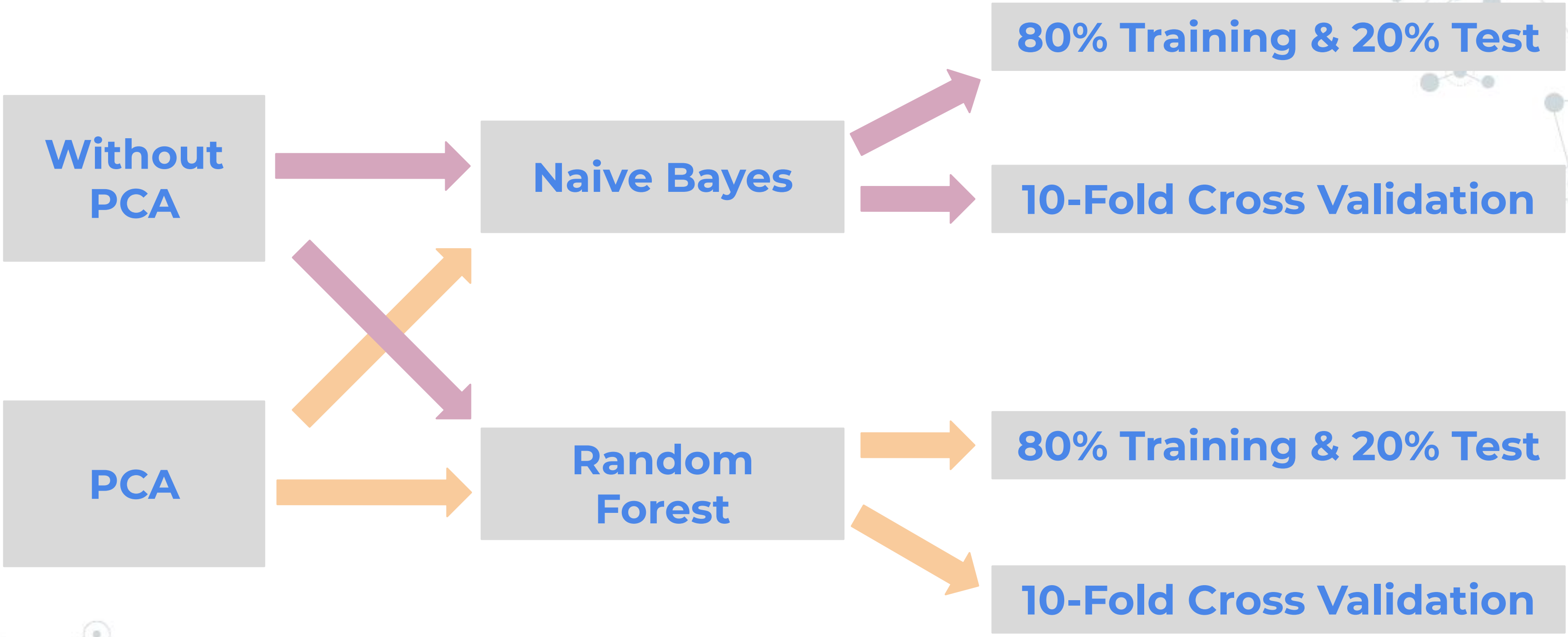
Louisiana	0.5045
South Carolina	0.4868
Maryland	0.4800
Florida	0.4583
North Carolina	0.4020
Alabama	0.3937
Georgia	0.3841
Delaware	0.3700
Kansas	0.3600
New Mexico	0.3450
California	0.3305

Data Visualization



State	City	Violent Crimes Per Pop
Alabama	Gadsdencity	0.9
Arkansas	Blythevillecity	0.9
California	EastPaloAltocity	0.97
	Comptoncity	0.9
	SantaFeSpringscity	0.86
	Inglewoodcity	0.86
Connecticut	Hartfordtown	0.94
	NewHaventown	0.88
Florida	Orlandocity	0.95
	LakeCitycity	0.87
	DaytonaBeachcity	0.86
Georgia	Brunswickcity	0.86
Indiana	Garycity	0.89
Maryland	Salisburycity	0.91
Massachusetts	Lawrencecity	0.88
Mississippi	Grenadacity	0.96
New Jersey	Bridgetoncity	0.93
	Trentoncity	0.85
	AsburyParkcity	0.85
New York	NewYorkcity	0.87
North Carolina	NewBerncity	0.91
	Fayettevillecity	0.86
Ohio	Limacity	0.97
	Youngstowncity	0.95

Models Overview



Without PCA

PCA

Naive Bayes

Random Forest

80% Training & 20% Test

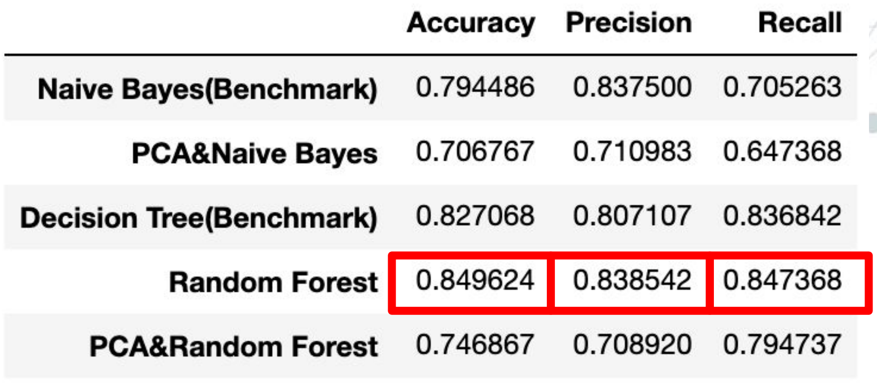
10-Fold Cross Validation

80% Training & 20% Test

10-Fold Cross Validation

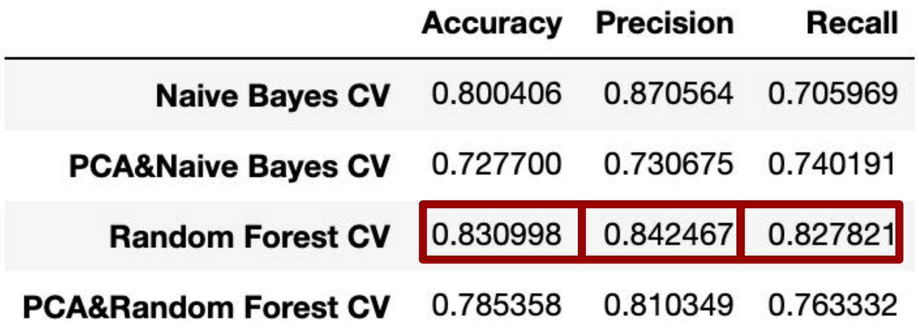
Models Overview

Training: 80%



	Accuracy	Precision	Recall
Naive Bayes(Benchmark)	0.794486	0.837500	0.705263
PCA&Naive Bayes	0.706767	0.710983	0.647368
Decision Tree(Benchmark)	0.827068	0.807107	0.836842
Random Forest	0.849624	0.838542	0.847368
PCA&Random Forest	0.746867	0.708920	0.794737

10-Fold Cross Validation

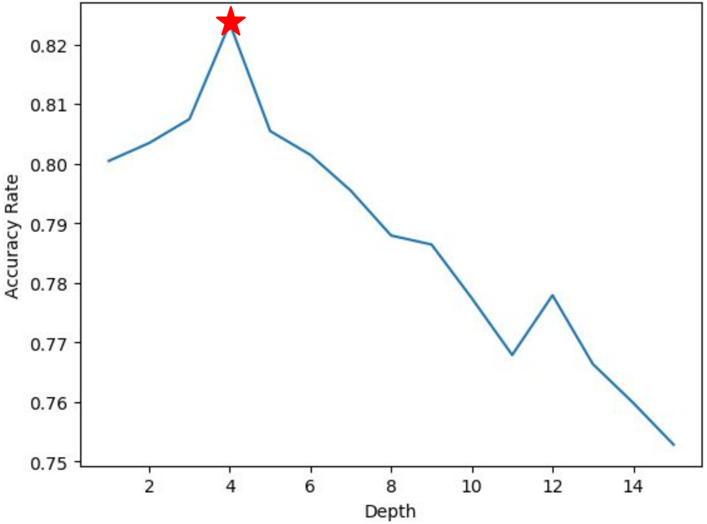


	Accuracy	Precision	Recall
Naive Bayes CV	0.800406	0.870564	0.705969
PCA&Naive Bayes CV	0.727700	0.730675	0.740191
Random Forest CV	0.830998	0.842467	0.827821
PCA&Random Forest CV	0.785358	0.810349	0.763332

Random Forest w/ 10-Fold Cross Validation

Optimal depth

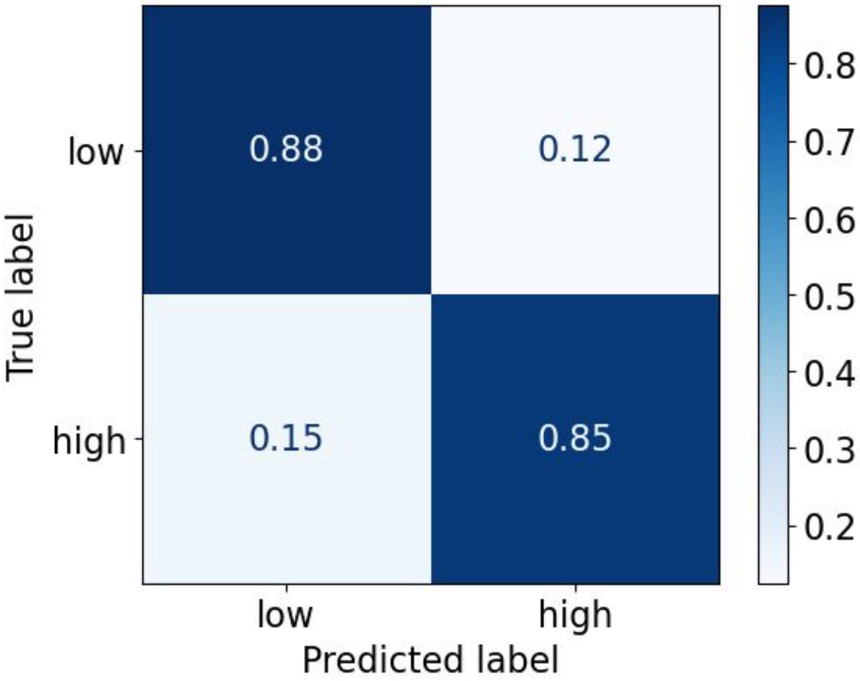
Decision Tree Model to pick optimal depth



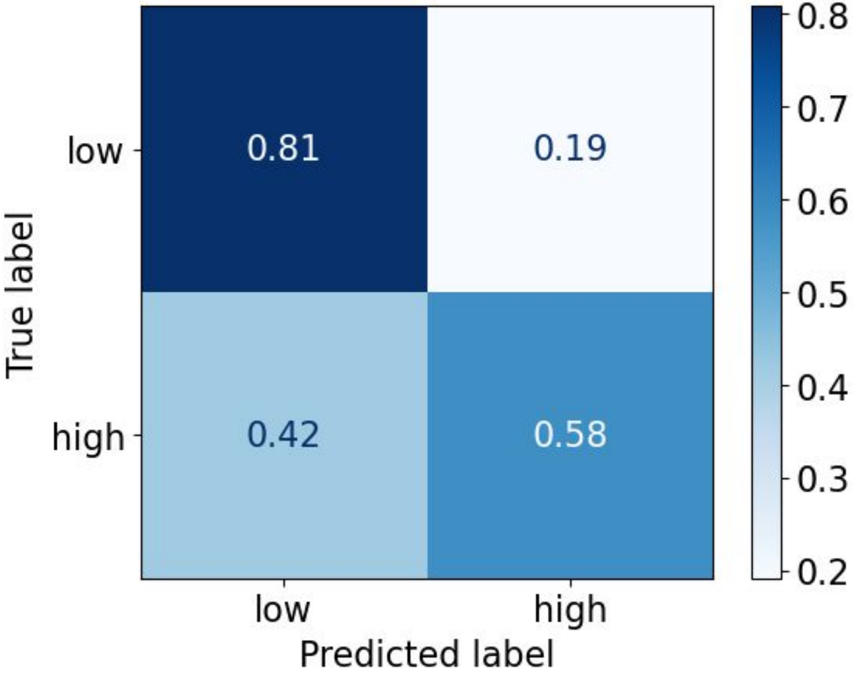
Optimal Depth = 4

Random Forest w/ 10-Fold Cross Validation

Random Forest only

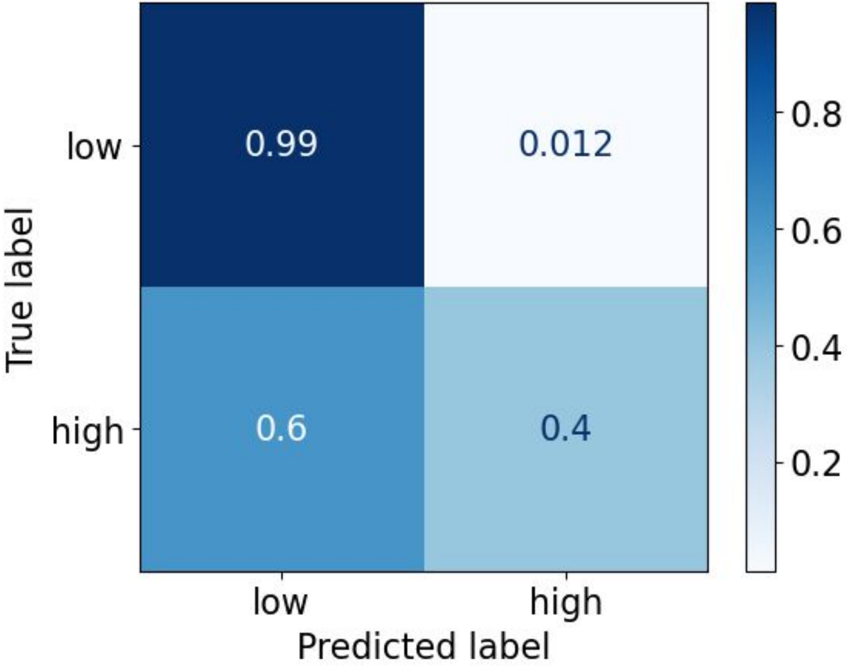


PCA & Random Forest

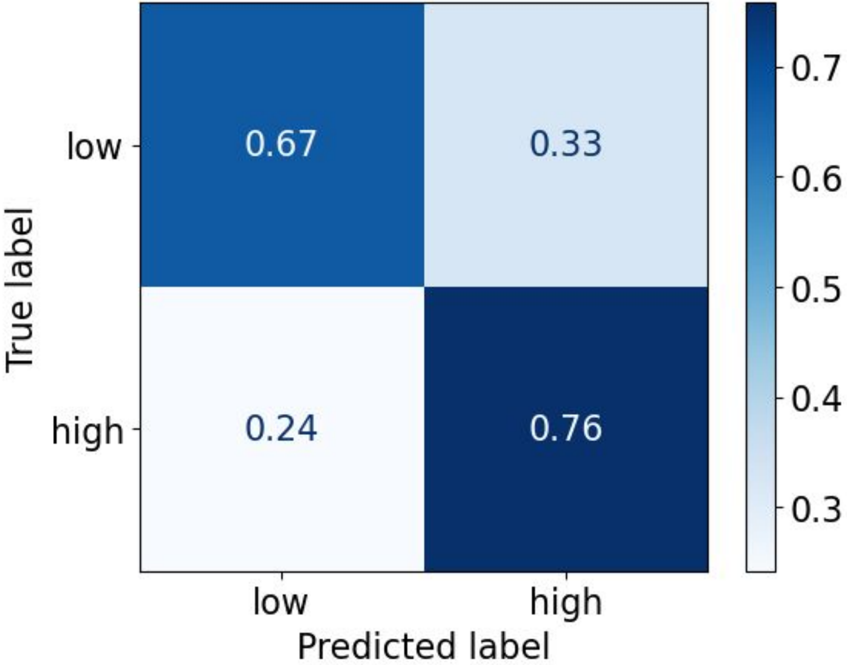


Naive Bayes w/ 10-Fold Cross Validation

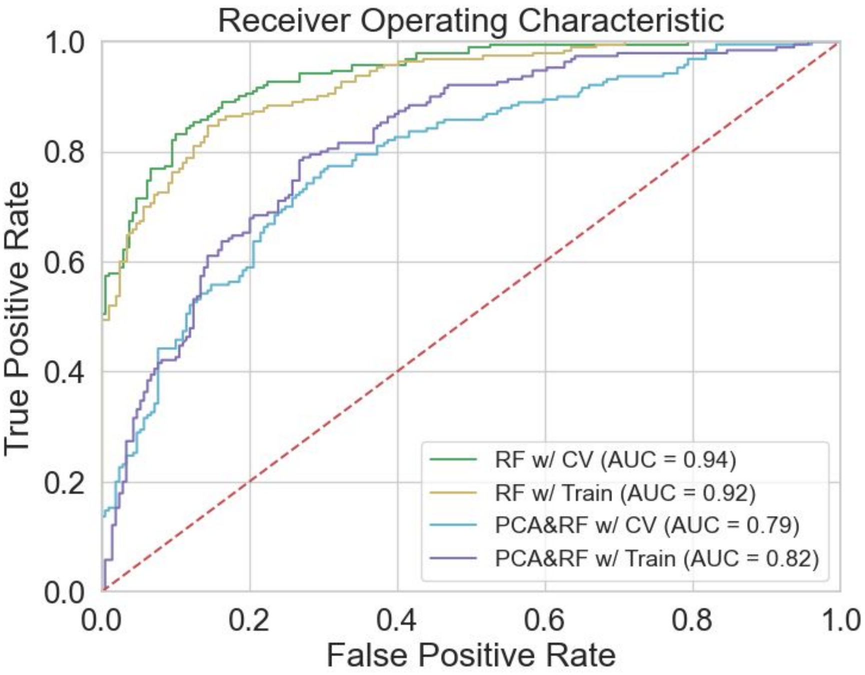
Naive Bayes only



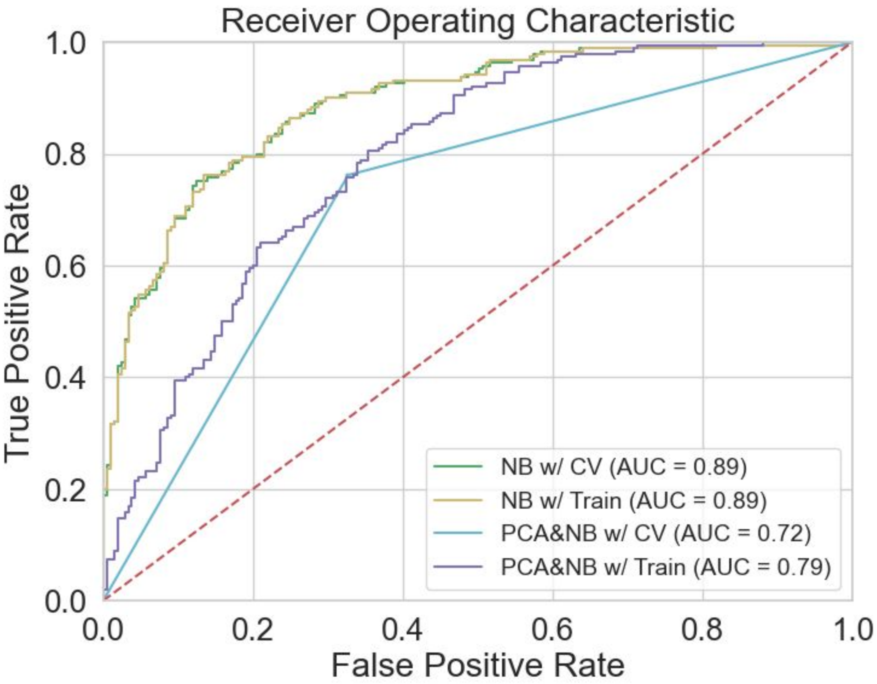
PCA & Naive Bayes



ROC Curve



Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
RF w/ CV (AUC = 0.94)
RF w/ Train (AUC = 0.92)
PCA&RF w/ CV (AUC = 0.79)
PCA&RF w/ Train (AUC = 0.82)



Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
NB w/ CV (AUC = 0.89)
NB w/ Train (AUC = 0.89)
PCA&NB w/ CV (AUC = 0.72)
PCA&NB w/ Train (AUC = 0.79)

Additional Due Diligence

- Explore attributes
- Incomplete dataset
  - Cities missing
  - Time not given
- Focusing modeling to one state

Thank you!

UCI University of California, Irvine
UCI University of California, Irvine