CASE

predict US crime rate

how to deal with a very small dataset?

radix

US CRIME RATE

Our Goals

Predict crime rate

Predict a crime rate for a given sample

Deal with a very small dataset

Understand some techniques to handle very small datasets

radix

TIMELINE

Case

EXPLORATORY DATA ANALYSIS

FEATURE ENGINEERING

HYPOTHESIS

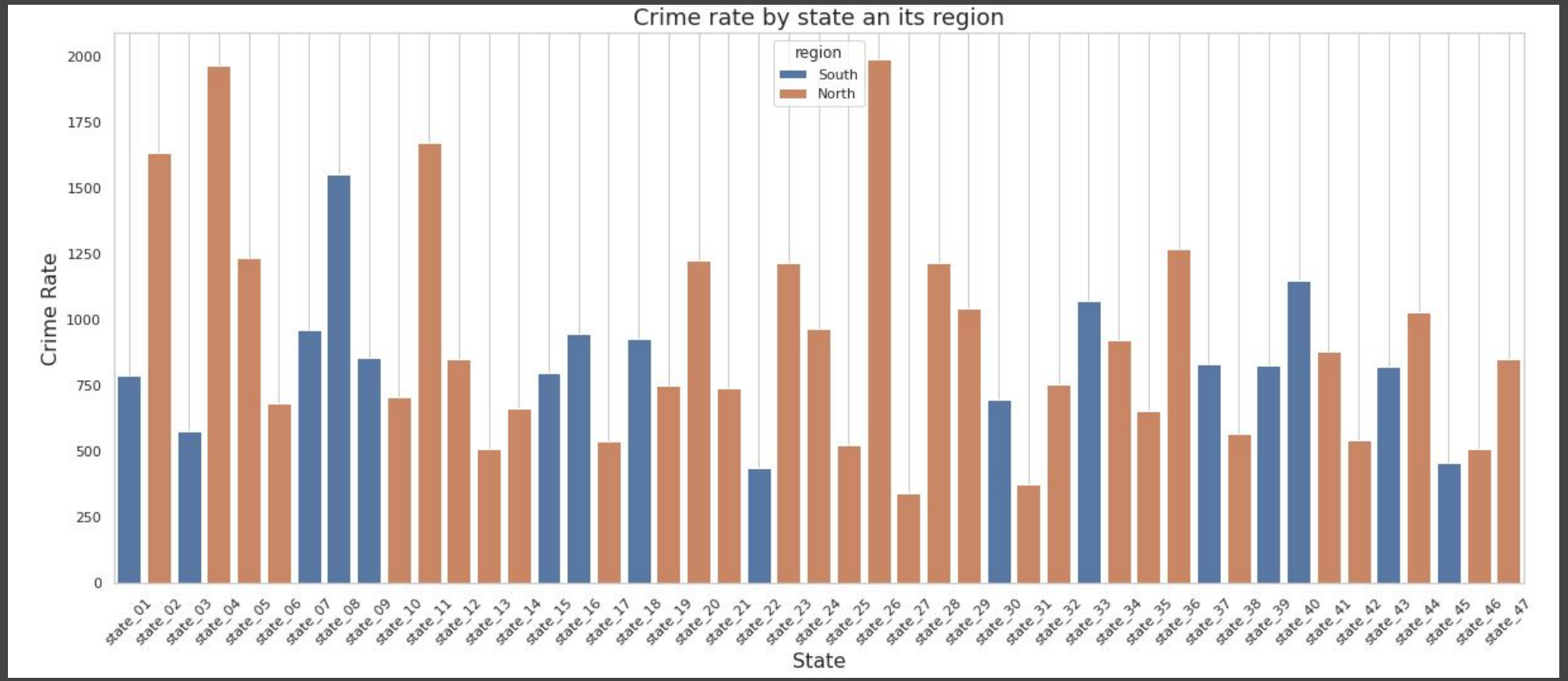
MODEL DEVELOPMENT

RESULTS

radix

#

What about the crime rate at 60's



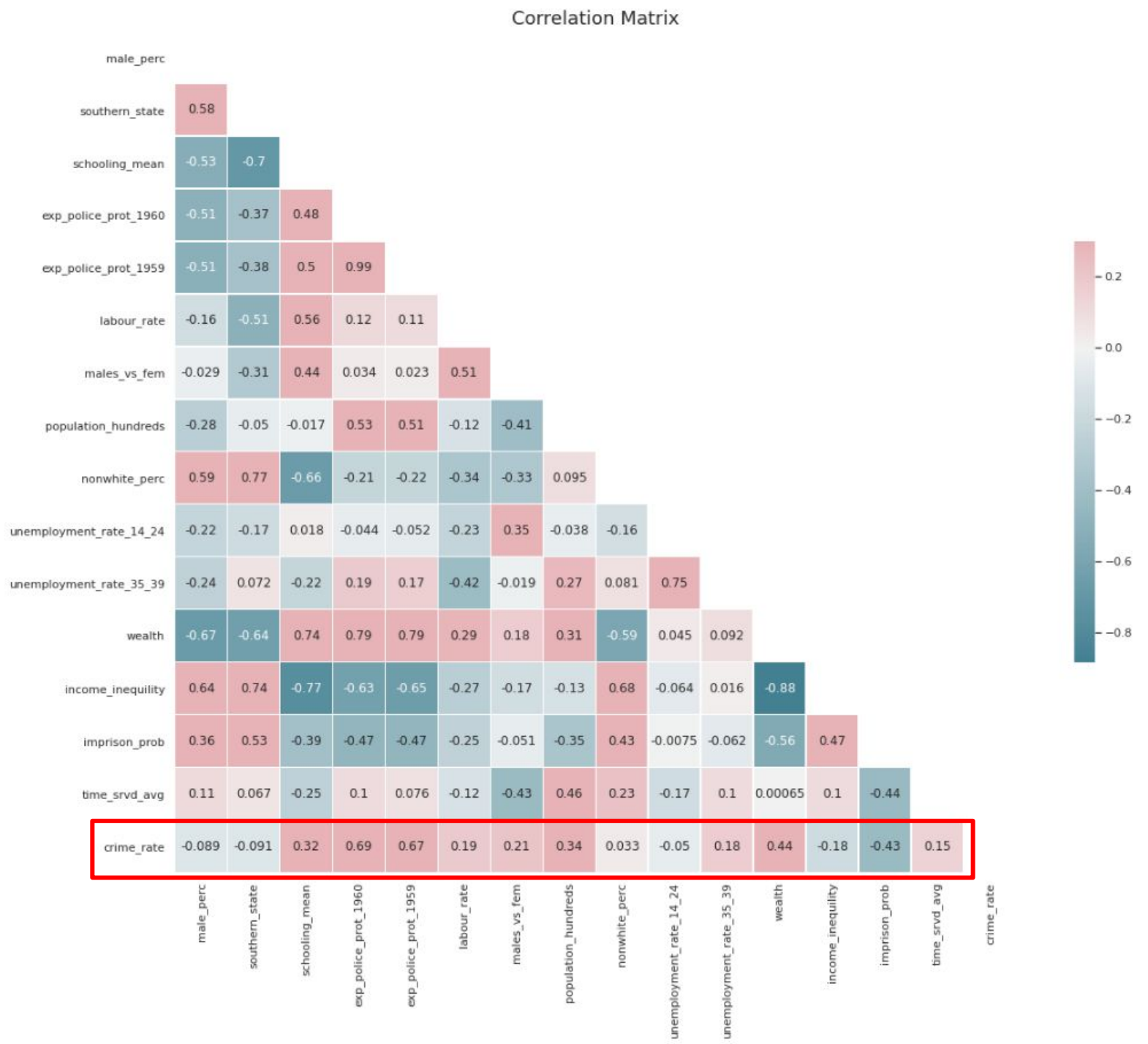
#

All features

Variable	Renamed	Description
M	male_perc	percentage of males aged 14–24 in total state population
So	southern_state	indicator variable for a southern state
Ed	schooling_mean	mean years of schooling of the population aged 25 years or over
Po1	exp_police_prot_1960	per capita expenditure on police protection in 1960
Po2	exp_police_prot_1959	per capita expenditure on police protection in 1959
LF	labour_rate	labour force participation rate of civilian urban males in the age-group 14-24
M.F	males_vs_fem	number of males per 100 females
Pop	population_hundreds	state population in 1960 in hundred thousands
NW	nonwhite_perc	percentage of nonwhites in the population
U1	unemployment_rate_14_24	unemployment rate of urban males 14–24
U2	unemployment_rate_35_39	unemployment rate of urban males 35–39
Wealth	wealth	wealth: median value of transferable assets or family income
Ineq	income_inequility	income inequality: percentage of families earning below half the median income
Prob	imprison_prob	probability of imprisonment: ratio of number of commitments to number of offenses
Time	time_srvd_avg	average time in months served by offenders in state prisons before their first release
Crime	crime_rate	crime rate: number of offenses per 100,000 population in 1960

#

Correlation



How to avoid overfitting?

Use simple models with barely none tuning

+ Be aware for the outliers
+ Select the features and avoid missing values and weak correlation
+ Combine models for the final answer.

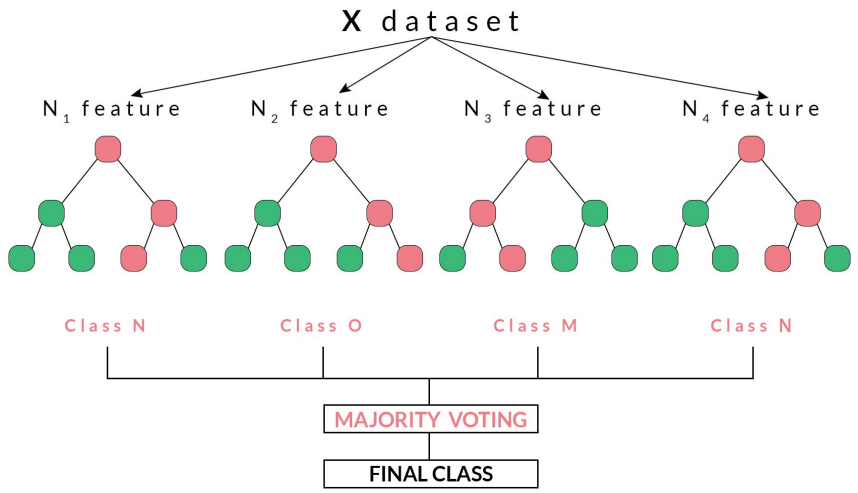
Hypothesis testing?

- Hypothesis 1
  - use All Features
- Hypothesis 2
  - schooling_mean exp_police_prot_1960 males_vs_fem population_hundreds wealth imprison_prob

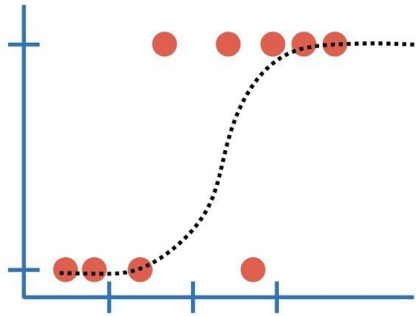
Models

- Random Forest
- Logistic Regression
- XGBoost

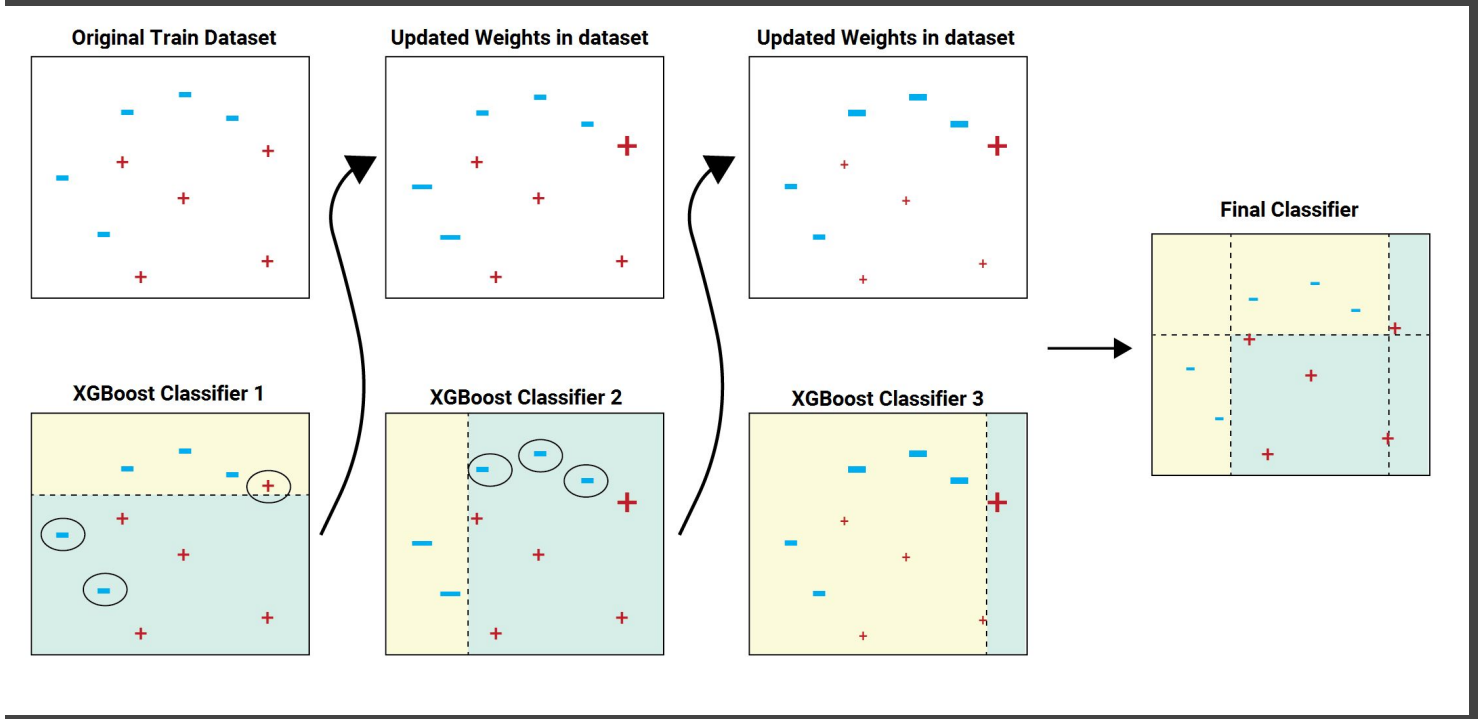
Models



Logistic Regression...



...Clearly Explained!!!



H1 RMSE 155.34 x 187.24 RMSE H2

radix

Question

Given the sample below predict the Observed Crime Rate.

male_perc = 14.0	labour_rate = 0.640	unemployment_rate_35_39 = 3.6
southern_state = 0	males_vs_fem = 94.0	wealth = 3200
schooling_mean = 10.0	population_hundreds = 150	income_inequility = 20.1
exp_police_prot_1960= 12.0	nonwhite_perc = 1.1	imprison_prob = 0.04
exp_police_prot_1959 = 15.5	unemployment_rate_14_24 = 0.120	time_srvd_avg = 39.0

Answer

The Predicted Crime Rate is 1099.85.
That's the number of offenses per 100000 population.
Since we have a population of 150 hundreds thousands.
So, we have a total of 164,977 offenses.

radix

Guilherme Lima

ADDRESS

84 Rua Porto da Capela
05345-010 São Paulo

CONTACT

guimarotto@gmail.com
+ 55 11 958896020

DATA SCIENTIST

Collaboratively empower customer experience with data-driven insights.

radix