Crime by County

Can we predict property crime based on other factors?

Reese Petersen

January 13, 2022

Overview

1. Introduction
2. Cleaning
3. Feature Selection
4. Model Selection
5. Testing
6. Conclusion

Introduction

Introduction

Predicting Crime in U.S. counties

Can we predict property crime levels based on other data?

Feature Categories:

- Weather
- Employment
- Income
- Demographics

Introduction
Data Sources

Websites:

- Crime: FBI: Offenses Known to Law Enforcement [3]
- Temperature & Precipitation: NOAA: Climate at a Glance [4]
- Income & Employment: BEA: Regional Data: GDP and Personal Income [2]
- Demographics: Census Bureau: County Population by Characteristics [1]

Cleaning

Cleaning
Crime

- State
- County
- **Property Crime***
- 69 (out of 2357) rows dropped for missing crime totals

* Crimes as reported by counties

Cleaning
Temperature & Precipitation

- State
- County
- Average Annual Temperature (°F)
- Total Annual Precipitation (in)

Cleaning
Income & Employment

- State
- County
- Per Capita Personal Income ($)
- Total Employment
  - Total Employment $\rightarrow$ Employment Rate
- Population

5 pairs of independent cities in Virginia combined with counties

Cleaning
Demographics

- State
- County
- Populations by age group, race, ethnicity.
  - Age groups are in 18 5-year bins up to 85+
  - 5 racial groups, 1 ethnicity, combinations
  - Race + Ethnicity groups dropped
  - Population $\rightarrow$ Fraction of Population

Warning!: Using race and ethnicity data leads to racially and ethnically biased models.

Feature Selection

Feature Selection

Highly Correlated Features

1. Highest absolute correlation between features
2. Feature with weaker target correlation removed
3. Repeat until highest correlation < 0.95

Feature Selection

Feature Sets

Table 1: Feature set sizes

Set	Features before filtering	Features after filtering
No Demographic Data	5	5
No Race, Ethnicity Data	61	42
All Data	327	218

Model Selection

Model Selection

Target Binning

PCR = Property Crime Rate, numbers are quantiles

- Low: PCR ≤ 0.33
- Moderate: $0.33 <$ PCR < 0.67
- High: ≥ 0.67

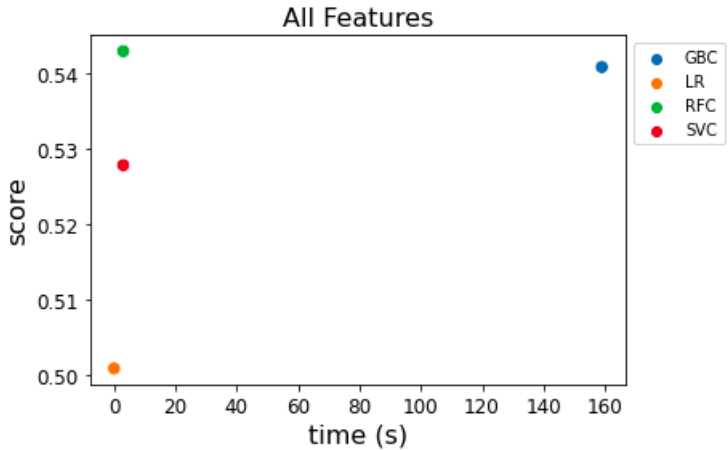
Model Selection

Classification Models

- Logistic Regression (LR)
- Random Forest Classifier (RFC)
- Gradient Boosting Classifier (GBC)
- Support Vector Classifier (SVC)

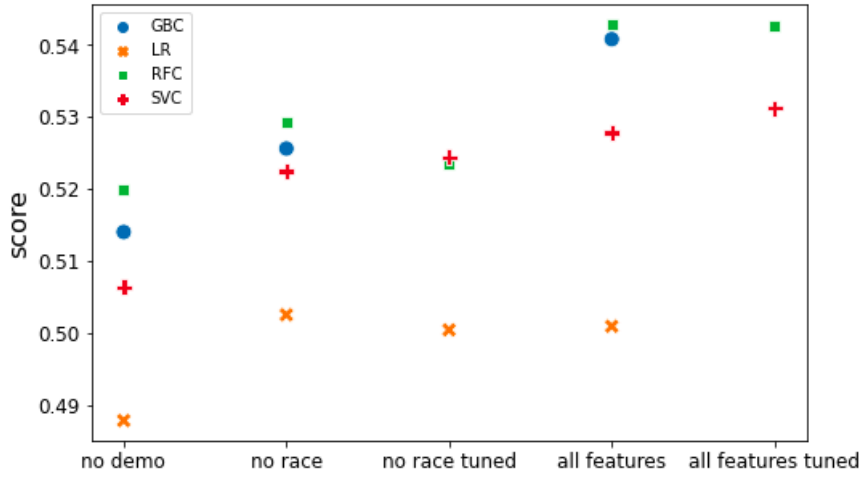
Model Selection

Score vs Time



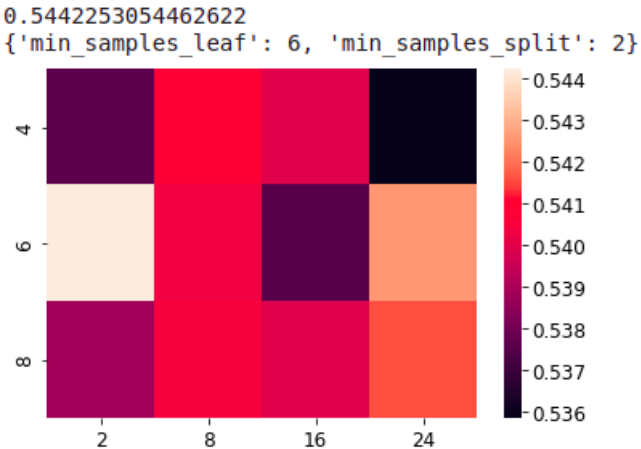
Model Selection

Model Scores



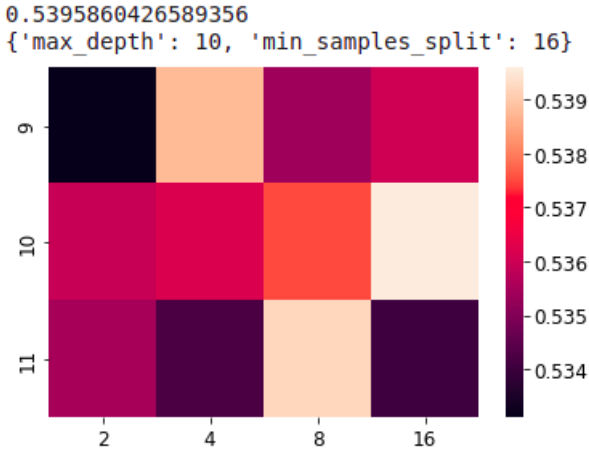
Model Selection

Hyperparameters, RFC



Model Selection

Hyperparameters, RFC



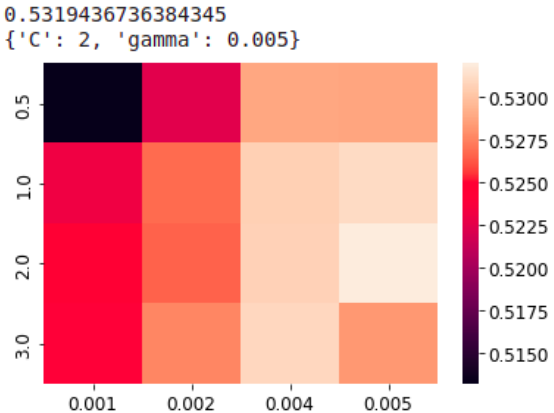
Model Selection

Hyperparameters, RFC

- best max depth: 10
- best minimum samples per leaf: 6
- minimum samples per split: 2
- 1000 estimators

Model Selection

Hyperparameters, SVC



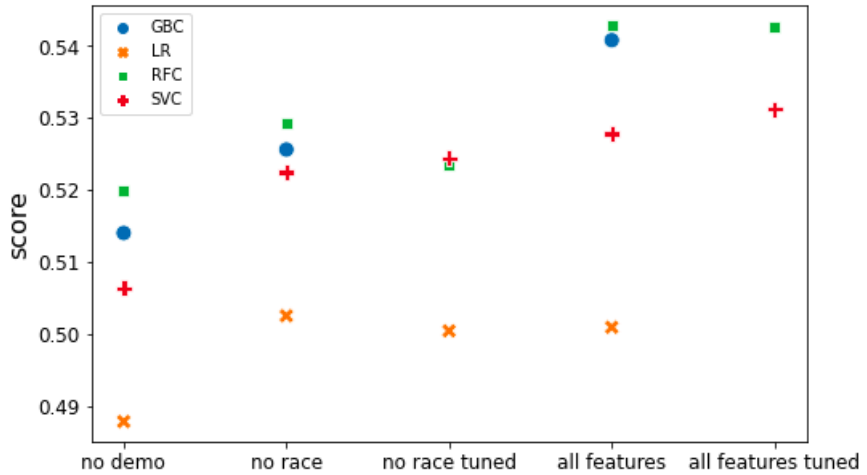
Model Selection

Hyperparameters, SVC

- best C: 2
- best gamma: 0.005

Model Selection

Model Scores, Revisited



Model Selection

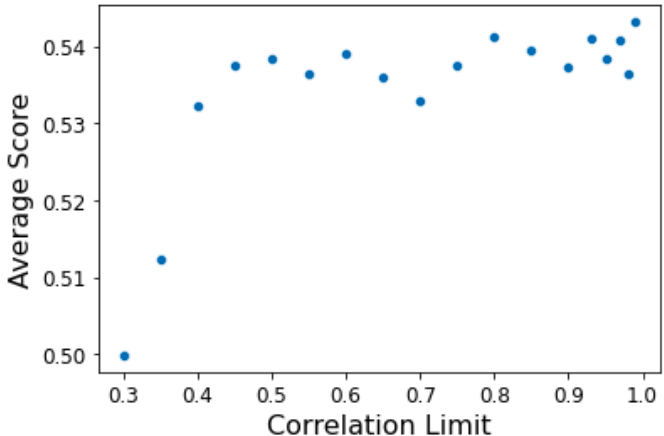
Correlation Limit, Revisited

What if we tune the feature selection instead?

1. Highest absolute correlation between features
2. Feature with weaker target correlation removed
3. Repeat until highest correlation $<$ Correlation Limit

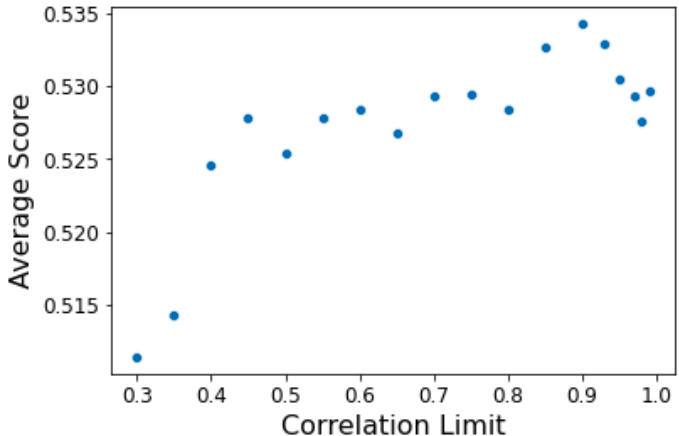
Model Selection

Score vs. Correlation Limit, RFC



Model Selection

Score vs. Correlation Limit, SVC

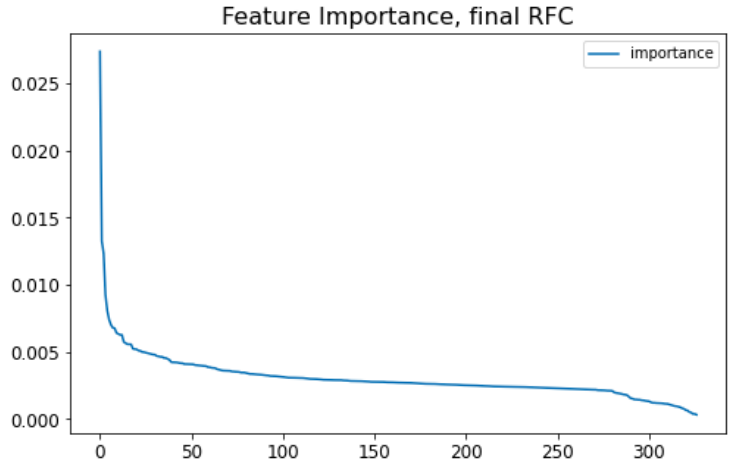


Testing

Testing
Final Model: RFC

- all features
- max depth: 10
- minimum samples per leaf: 6
- cross-validation R^2 score: 0.54
- accuracy on 2019 data: 0.66 (no race/ethnicity data: 0.61)

Testing
Feature Importance

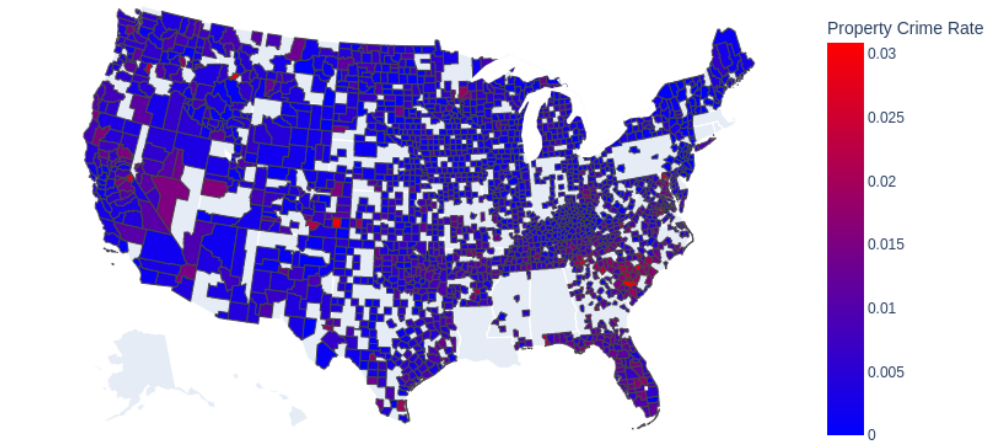


Testing
Feature Importance

	feature	importance
0	Annual Average Temp	0.027386
1	Total Annual Precipitation	0.013268
2	Employment Rate	0.012342
3	FRAC_WA_FEMALE_5	0.009249
4	FRAC_TOM_FEMALE_14	0.008064
5	FRAC_TOT_FEMALE_5	0.007399
6	FRAC_TOT_FEMALE_15	0.007018
7	Per capita personal income ($)	0.006794
8	FRAC_TOM_FEMALE_16	0.006767
9	FRAC_TOM_MALE_16	0.006419

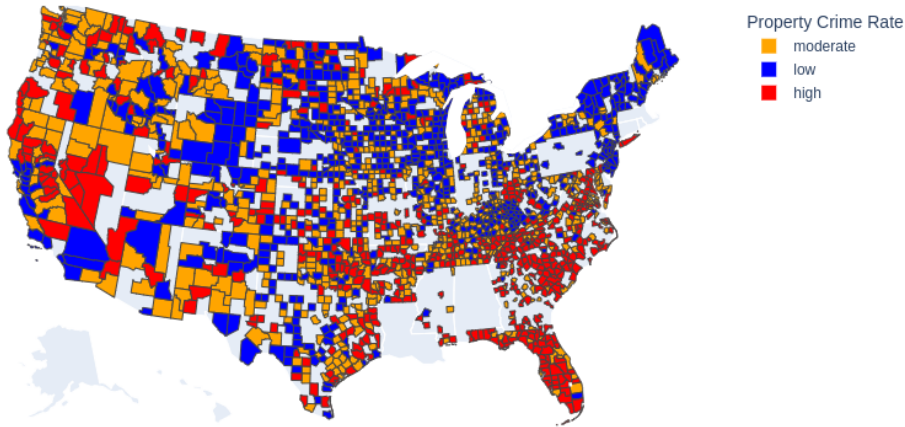
Testing

Property Crime Rate 2019



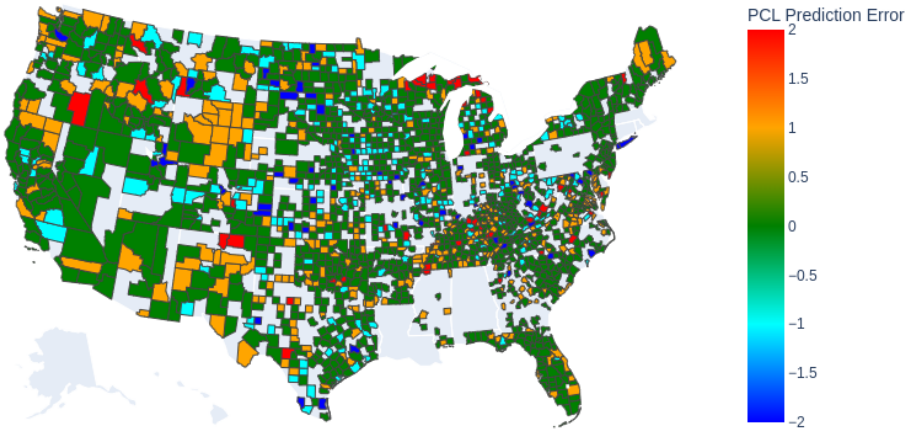
Testing

Property Crime Level 2019



Testing

Property Crime Prediction Error 2019



Conclusion

Conclusion
Uses

- Generally: Estimate property crime level
- In Data Science: Study bias in machine learning models

Conclusion
Strengths

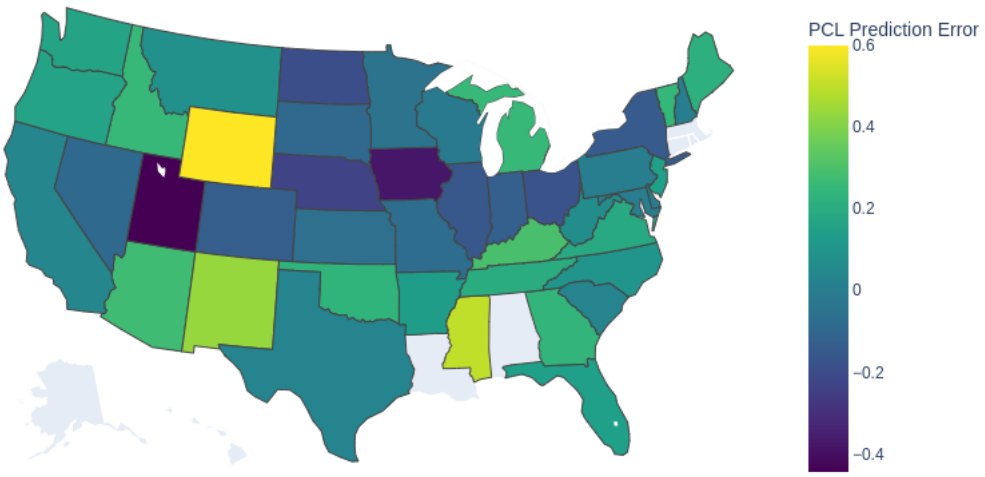
- More granular than Metropolitan Statistical Areas
- Features are easy to get
- Nationally applicable
- Bias has low impact, easy to remove

Conclusion
Weaknesses

- Crime numbers aren’t totals: city agencies
- Not granular enough
- About 24% of counties don’t report
- Racially and ethnically biased

Questions?

PCL Error by State



Backup slides

[plain,noframenumbering]

References

[1] United States Census Bureau. *County Population by Characteristics*. URL: https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html. (accessed 12.20.2021).

[2] US Department of Commerce Bureau of Economic Analysis. *GDP and Personal Income*. URL: https://apps.bea.gov/iTable/iTable.cfm?reqid=70&step=1&acrdn=6. (accessed 12.20.2021).

[3] Federal Bureau of Investigations. *Offenses Known to Law Enforcement*. URL: https://ucr.fbi.gov/crime-in-the-u.s/2018/crime-in-the-u.s.-2018/tables/table-10/table-10.xls/view. (accessed 12.21.2021).

[4] National Oceanic National Centers for Environmental Information and Atmospheric Administration. *Climate at a Glance*. URL: https://www.ncdc.noaa.gov/cag/county/mapping/110/tavg/202012/12/mean. (accessed 12.20.2021).