Investigating County-level Theft Rates in the U.S. in 2020

STAT 471 Final Project

Mason Shihab
Diyang Chu

December 19, 2021

Executive summary

Problem

We aim to generate insights as to features that might predict variation in the rates of thefts committed per capita. In doing so, we hoped to understand if economic, health, criminal justice, or other types of variables could be used to help with this analysis.

Data

We collected data from a variety of mostly government sources, with a plurality coming from the American Community survey in particular, and its parent organization, the US Census Bureau, in general. Our response variable, "theft crimes known to law enforcement," was provided by the FBI. Other variables were aggregated by nonprofit organizations such as The Marshall project and the Brookings Institute. Also of note is the Police Scorecard Project. Comprehensive information about our data sources is available in the Data section below.

Analysis

We performed our analysis using two distinct supervised learning methodologies: penalized regression as well as tree based methods. We use both of these tools to create predictive models. However, rather than attempting to predict future thefts, we instead use these analyses to look backwards and attempt to gain insights and associations with thefts committed in 2020. After reviewing our models to gain insights, we compare them against each other as well as the intercept-only model via root mean squared error metric so that the error rate can be on the same scale as the response variable.

Conclusions

Our models are difficult to draw any conclusions from because they are not significantly more effective than the intercept-only model. Nevertheless, we do find some interesting associations using the gradient boosting model. The prime takeaway here is that the rate of theft cannot be easily predicted, even using metrics that would likely be seen as predictive, such as police accountability, poverty, and health metrics. Indeed, the biggest conclusion may be that there is no satisfactory conclusion here. Current data may not be sufficient for a proper national analysis, and future research may need to be focused on the local level, with partnership from governments to provide complete data.

Introduction

The context of the problem that we chose to analyze is twofold. The first has to do with the fact that there is a growing movement behind instituting new Criminal Justice Reform measures. This moment, and the policies it endorses, seeks to reimagine our justice system in a way that reduces crime and keeps people out of prison at the same time. Certainly, this is a worthy aim that should be investigated. In addition to that, we are entering into a period of growing economic inequality and turmoil. These uncertain conditions contribute to a growing debate as to the scope, size and magnitude of government in general and its public assistance programs in particular. And in the context of the covid-19

pandemic, social safety nets such as unemployment insurance came strongly into public view and consideration.

This context provides a high degree of uncertainty that calls for equally high levels of quantitative analysis. This is why we sought a single response variable that could capture these potential factors (and others) to provide insight into their relative influence on that chosen variable. This variable would need to be universally well-documented and sensitive to local on-the-ground differences among communities. We decided, then, that the most interesting variable for us to analyze is the amount of theft crimes per capita. In our view, theft is an economic crime that is potentially impacted both by traditional legal and criminological indicators as well as financial and even political factors. This makes it particularly interesting as a response variable in a multi-faceted analysis.

To that end, our primary goal of this analysis consists in identifying the most predictive and impactful factors that are associated with theft. Our variables fall into five primary buckets: demographic, socioeconomic, strength of social safety net, criminal justice policy response, and health-related factors. Within each of these categories, we have included several unique ways of measuring that category's impact via the variables within it. In general, a successful supervised learning analysis, in our view, is one which gives some indication that some variables–and not others–are highly impactful for predicting its response. However, our results are not strong enough to make any significant claims. However, we do not consider this project a failure, as it shows that publicly available data on this topic is not yet strong enough to build highly predictive models on the national level.

We performed this analysis using state-of-the-art machine learning predictive techniques that in turn automatically select the most important variables for predicting response variables. This kind of insight can be used to support policymakers who wish to bring down rates of criminal activities by providing them direction on where to focus public resources. There are many theories and schools of thought which hope to explain not only why crime occurs, but how to get rid of it. We feel that the breadth of our chosen variables gave each of these theories an equal chance at obtaining new evidence for their cause. We hope our findings will be used to direct new research on this field and demonstrate that better, more comprehensive data is needed on this topic.

Data

We included a large variety of data sources and methodologies for cleaning and analyzing this data. This can all be seen in our repository: https://github.com/masonshihab/theft-crime-2020

Data sources

American Community Survey: The survey is meant to be a guide for policymakers in directing how certain public funds should be spent. The Census Bureau contacts more than 3 million households per year for this survey. It is one of the most comprehensive publicly available datasets available. We Downloaded the data directly into R from the census bureau using the Tidycensus package. In order for our data to be reproducible from the top, other users will need to install their own US Census API keys, which are provided at the following website: https://api.census.gov/data/key_signup.html. Once a key is

obtained, this code will need to be ran: *census_api_key(key, overwrite = FALSE, install = TRUE)*. More information can be found at https://rdrr.io/cran/tidycensus/man/census_api_key.html.

The land area dataset contains the square mileage of every county in the nation and was put together by the Statistical Compendia Program, which included the USA counties database (which in turn contains the land area dataset). This information was used to calculate the population and housing density. The raw form of this dataset is available at https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html#LND.

The 2020 County Health Rankings dataset was obtained from the website for The County Health Rankings & Roadmaps Program. A team at the University of Wisconsin Population Health Institute collected many county-level measures from a variety of national and state data sources. The dataset includes health situations of nearly all counties in the U.S. According to the website, these measures used in this dataset were standardized and combined using scientifically-informed weights. The dataset can be found here: https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation.

We obtained the 2018 Social Vulnerability Index dataset from Data.gov, which is a website managed and hosted by the U.S. General Services Administration, Technology Transformation Service. The Agency for Toxic Substances and Disease Registry's Geospatial Research, Analysis & Services Program compiled relevant data from the U.S. Census to identify the communities that will most likely need support for public health and emergency response. Though our focus is the year 2020, we had to use the 2018 dataset because it is the newest available one. However, we believe that the 2018 data will provide reliable information as the years are close enough. The original dataset is available here: https://catalog.data.gov/dataset/social-vulnerability-index-2018-united-states-county.

Small Area Income and Poverty Estimates (SAIPE) Program from the US Census: This data set consists of estimates by the Census Bureau that is a combination of surveys and other information including administrative records. Data downloaded and manually cleaned in part in Excel to remove styling before loading into R for additional cleaning and merging. Data can be found here: https://www.census.gov/data/datasets/2018/demo/saipe/2018-state-and-county.html.

State and Local Direct General Expenditures, Per Capita: these data were compiled by the tax policy Center at the urban Institute and Brookings Institution. Specifically, they were compiled by the Urban Institute via the project “State and local Finance data: exploring the census of governments.” The data were obtained through the US Census Bureau's “Annual Survey of State and Local Government Finances. In other words, the original source of the data were the local governments themselves. Data can be found at: https://www.taxpolicycenter.org/statistics/state-and-local-direct-general-expenditures-capita

SOI Tax Stats - Individual Income Tax Statistics - 2018 ZIP Code Data (SOI) through the IRS's publicly available zip level data. The raw counts were obtained by the IRS's reporting of various tax and benefit claims through official forms that taxpayers must complete every year, such as the 1040. Data downloaded and manually cleaned in part in Excel to remove styling before loading into R for additional cleaning and merging. Data can be found at: https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2018-zip-code-data-soi

The Adults in Correctional Facilities dataset was obtained from observablehq.com. The Marshall Project extracted the number of adults in correctional facilities per county from the 2020 Decennial Census. Here, "correctional facilities" refer to federal detention centers (e.g., immigration detention centers), federal and state prisons, local jails, correctional residential facilities (e.g., halfway houses), and military jails. The raw dataset can be found here: https://observablehq.com/@themarshallproject/adults-in-correctional-facilities-from-decennial-census.

We found the Police Scorecard dataset on the website for The Police Scorecard Project, which was founded by Samuel Sinyangwe and a team of data scientists. It is the first nationwide public evaluation of policing in the U.S., which reports levels of police violence, accountability, and other policing outcomes for over 16,000 municipal and county law enforcement agencies. The team collected data from a variety of sources, including the FBI Uniform Crime Report, the Bureau of Justice Statistics' Annual Survey of Jails, the U.S. Census Bureau's Survey of State and Local Government Finances, the California Department of Justice's OpenJustice database, etc. The Scorecard dataset was published online in early 2021, but the data used to calculate the scores were collected in different years, ranging from 2012 to 2020. The datasets can be found here: https://policescorecard.org/.

The COVID-19 Community Vulnerability Crosswalk dataset was found on HealthData.gov, a federal government website managed by the U.S. Department of Health & Human Services (HHS). One spreadsheet in the Crosswalk dataset contains the evaluation of the extent to which each county in the U.S. was severely affected by the COVID-19 pandemic. HHS collected the data from HSS agencies and HHS's state partners, including the Centers for Medicare and Medicaid Services, Centers for Disease Control and Prevention, Food and Drug Administration, and the Agency for Health Care Research and Quality, etc. The dataset was published in May 2021. The original dataset is available here: https://healthdata.gov/Health/COVID-19-Community-Vulnerability-Crosswalk-Crosswa/x2y5-9muu.

We obtained the Atlas of Rural and Small-Town America dataset from the website of Economic Research Service (ERS), which is a component of the United States Department of Agriculture. The dataset provides information based on five major categories of socioeconomic factors, including people, jobs, county classifications, income, and veterans. ERS integrated data from the American Community Survey (ACS), the Small Area Income and Poverty Estimates (SAIPE), and the Bureau of Labor Statistics and other sources. The dataset was updated in June 2021. The original dataset can be obtained here: https://www.ers.usda.gov/data-products/atlas-of-rural-and-small-town-america/download-the-data/.

2020 US Presidential Results: These are the election results by county. After the election, these data were quickly made available on state government websites. They were then compiled by the MIT Election Data and Science Lab (MEDSL), in partnership with the Harvard Dataverse into this single source. The data was downloaded into the local machine first because we could not locate a raw CSV. Rather, the data is available as part of a zip file. A link to the project can be found here: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ

Data cleaning

We began our project by first identifying a suitable response variable. We were able to go to the FBI's website and find crime statistics with their publicly-available crime data Explorer.[1] Within this data Explorer we found annual reports for the year 2020, under the Offenses Known to Law Enforcement tab. By clicking to access this data, the user is directed to download a zip file which contains all the offenses known to law enforcement that have been compiled by the FBI through partnership with county and city police agencies throughout the nation. In order to combine cities with counties, we used a publicly available data set which matches zip codes, counties, and cities to then convert the city data to county data. We then merged two these two data sets to create a master 2020 "known crime" dataset. Within this, we aggregated the three primary types of theft tracked by the FBI: burglary, larceny and theft, and motor vehicle theft. We later divide this by total population to derive the "theft rate."

The next step was to find additional outside data that could predict changes in theft rate by county, which are detailed above. Though some of our data sets required more cleaning than others, in the end we were able to compile them into individual observation rows at the county level. We then merged these data sets either by the combination of state and county or by the fips code using the inner join command. We used a combination of traditional tools found in the Tidyverse, as well as more advanced libraries such as the "tm" and "usdata" packages to create a true "tidy" dataset, with one row for each county.[2] Several of our features were missing a few counties, including a feature we felt would be very important to include. This was the incarceration rate by county. However, this feature was missing several hundred counties. Rather than exclude those counties, we chose to impute these missing values using a simple random Forest. The code for this can be found in section "2.1-imputation.r" in our repository. Overall, we were able to retain 2293 counties in our final analysis, with 65 features obtained from 12 publicly available data sets.

Data description

There are 2293 observations in the data. Each observation represents a county in the US. Our response variable is county-level theft rate in 2020, which is a continuous variable. There are 65 features in the data, all of which are numerical. We categorized the features into five broad categories, which are Basic demographics, Socioeconomic status (SES), Social safety net, Criminal justice response, and health-related factors. There are 25, 25, 5, 5, or 5 features, respectively, in each category. For a detailed description of these features, refer to the Appendix.

Data allocation

We split our dataset into two subsets using an 80-20 split. The training dataset consists of 80% of all observations, and the testing dataset consists of 20% of all observations. The training dataset was used for data exploration and building predictive models. The testing dataset was used for evaluating our models. Before splitting, we utilized a random seed to ensure reproducibility. After splitting, we saved the

[1] This system can be accessed at https://crime-data-explorer.app.cloud.gov/pages/explorer/crime/crime-trend.
[2] https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html

two subsets into two separate documents, which helped avoid repeatedly conducting the train-test split for each class of methods.

Data exploration

EDA for Response

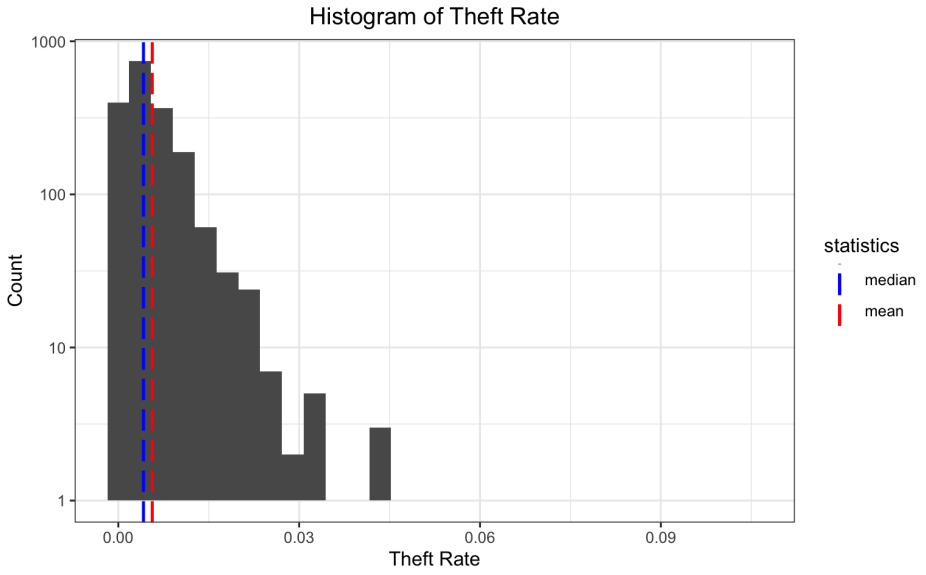

Figure 1: This is a histogram of the response variable in the data.

To understand the distribution of the response variable, we first plotted a histogram of theft rate. As seen from Figure **1**, the data appears to be right-skewed, with some counties exceeding a theft rate of 4%. The mean county-level theft rate in 2020 is 0.563%; the median is 0.417%. There are a few outlier counties with very high theft rates.

State	County	Theft Rate
Indiana	Hamilton	0.10498
New York	New York	0.07927
Colorado	Jefferson	0.07638
Colorado	Denver	0.04860
Mississippi	Tunica	0.04465
California	San Francisco	0.04427
Missouri	Greene	0.04167
West Virginia	Wayne	0.04136
Missouri	Marion	0.03510
Georgia	Bibb	0.03394

Table 1: This is a table of the top 10 counties with the highest theft rate.

We proceeded to determine which counties had extremely high theft rates in 2020 by looking at the sorted data. The sorted data in Table **1** shows that the top 6 counties with the highest rates of theft incidence are

Hamilton county in IN, New York county in NY, Jefferson county in CO, Denver county in CO, Tunica county in MS, and San Francisco county in CA.

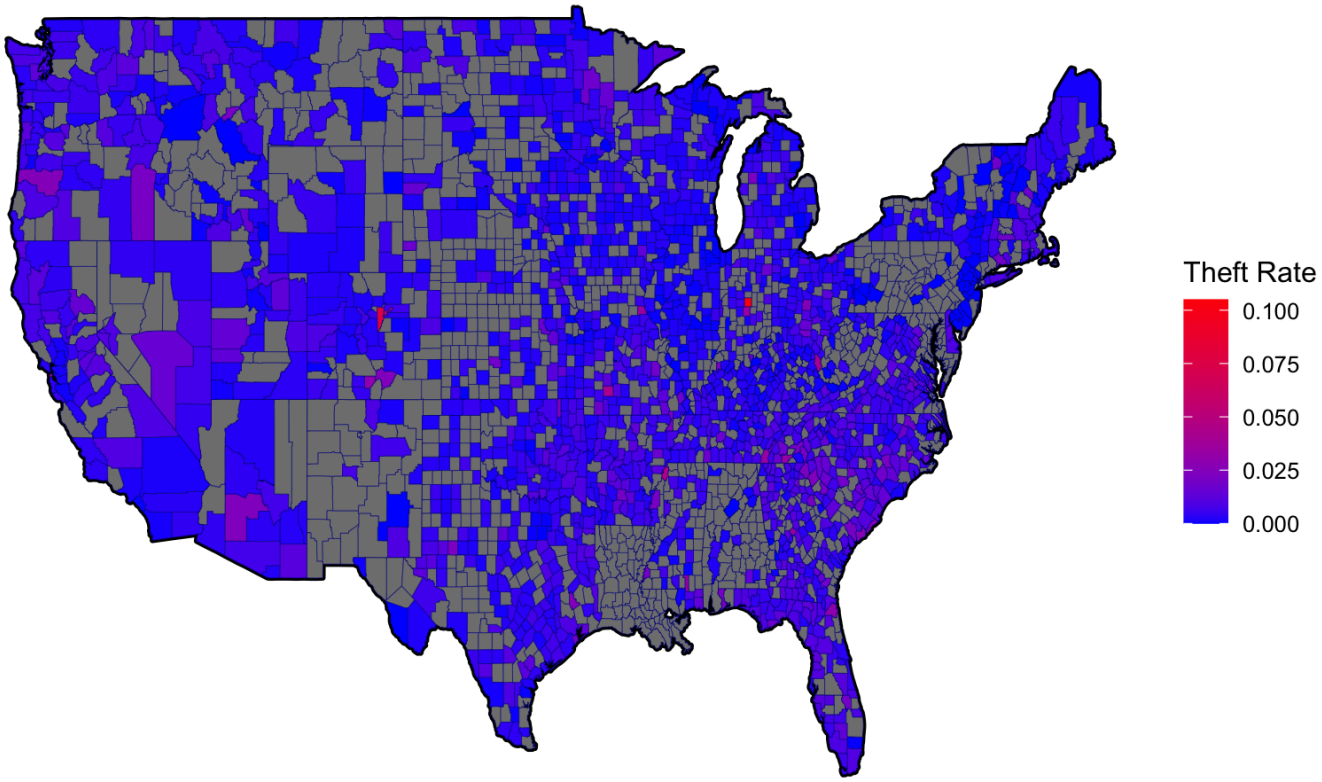

Figure 2: Heap map for the response variable in the data.

We also created a heat map for our response variable. As shown in Figure **2**, most of the counties included in this dataset have theft rates well below 2.5%. Very few states have theft rates of about 7.5%. The grey areas indicate counties not included in our dataset.

EDA for Features

Correlation Plot for Social Safety Net

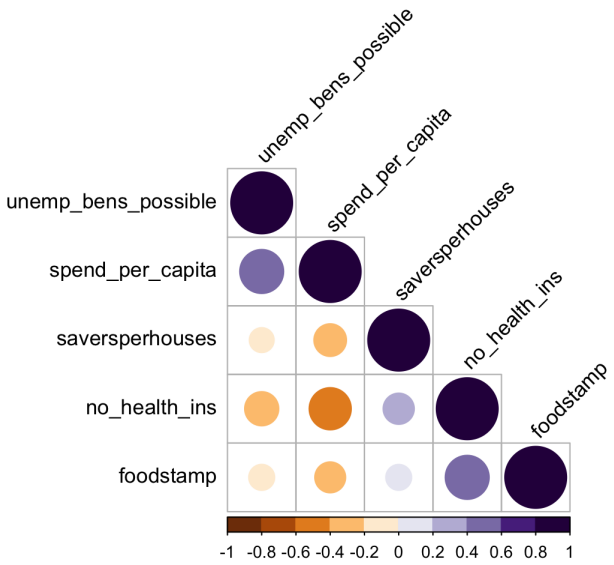

Figure 3: A correlation plot for the 5 features in the category of Social Safety Net

In Figure **3**, we observed a positive correlation between State and local government spending on people and State unemployment insurance. Also, Percent of households qualifying for food stamps is positively correlated with No Health Insurance. Not surprisingly, No Health Insurance is negatively correlated with State and local government spending on people.

Correlation Plot for Criminal Justice Response

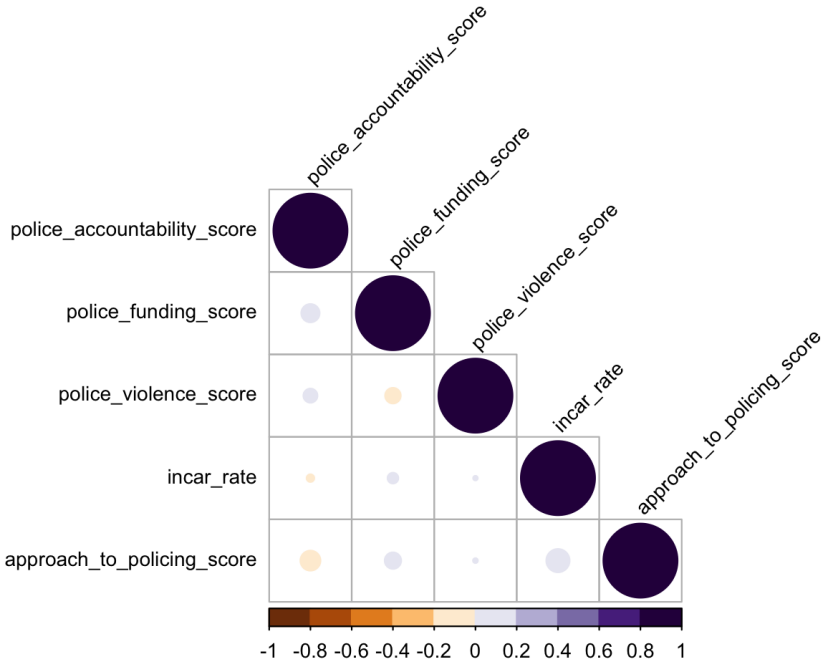

Figure 4: A correlation plot for the 5 features in the category of Criminal Justice Response

In Figure **4**, we found no significant correlations among features belonging to the category of Criminal justice response.

Correlation Plot for Health-related Factors

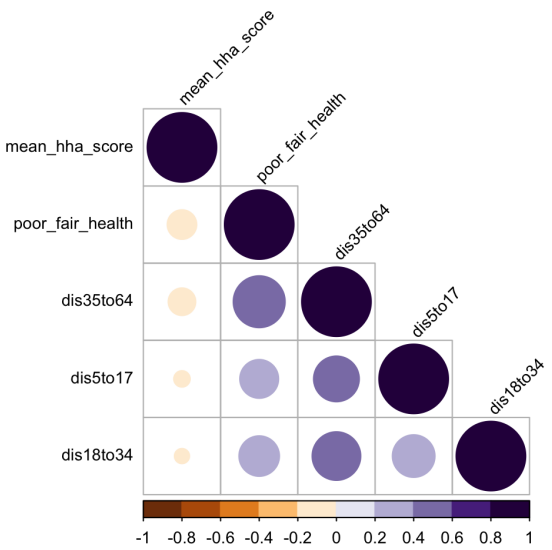

Figure 5: A correlation plot for the 5 features in the category of Health-related Factors

In Figure **5**, we observed positive correlations between Percent adults reporting poor or fair health and Percent of people with disability in each of the three age groups. Moreover, the percentages of people with disability for the three age groups are also positively correlated with each other.

Correlation Plot for SES

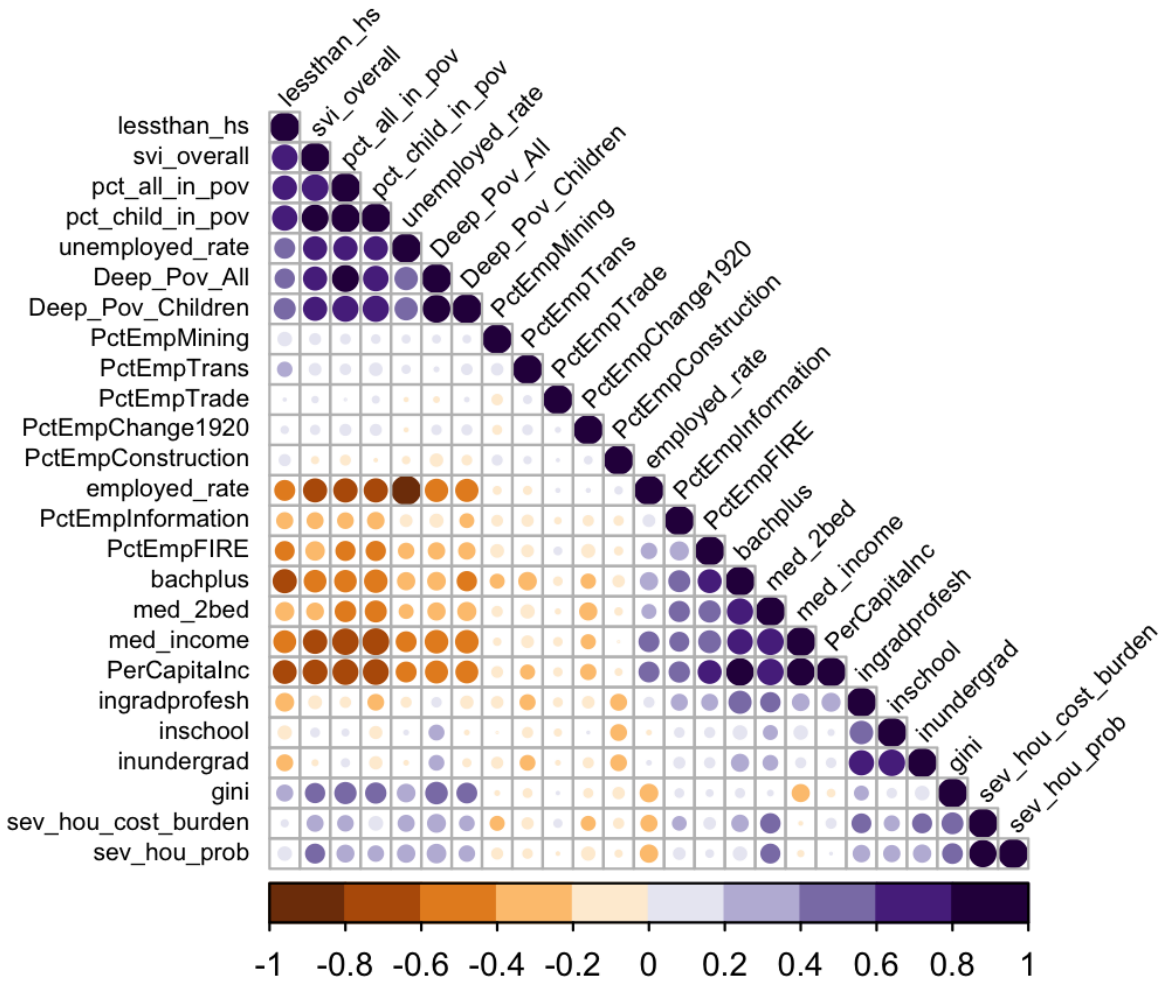

Figure 6: A correlation plot for the 25 features in the category of SES

In Figure **6**, we observed that there are positive correlations between the Social Vulnerability Index (SVI) and all poverty-related features. We also found that a group of features, including Employment rate, Percent with college or higher education, Median household income, and Per capita income in the past 12 months, are negatively correlated with SVI and poverty-related features.

Correlation Plot for Basic Demographics

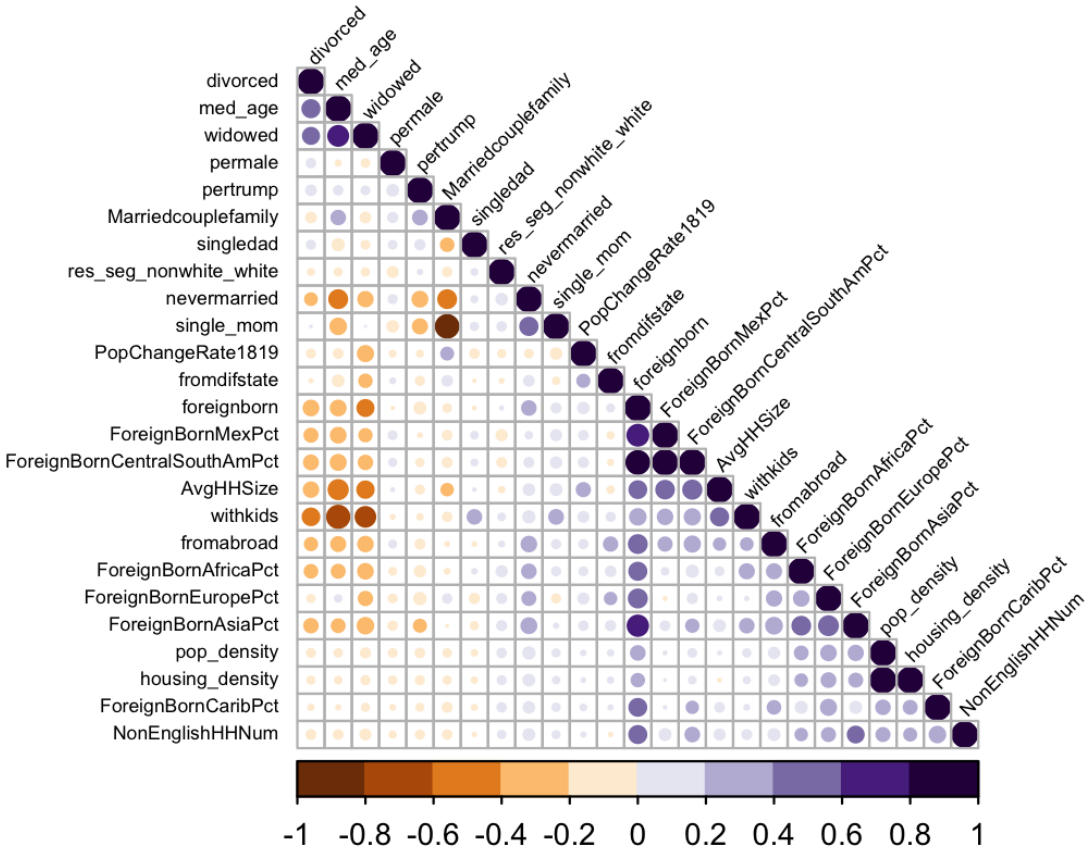

Figure 7: A correlation plot for the 25 features in the category of Basic Demographics

From Figure **7**, we observed that Percent of household with own children is negatively correlated with both Age (median) and Percent Divorced. Housing density has a significant positive correlation with Population density.

Correlation Plot for All Features

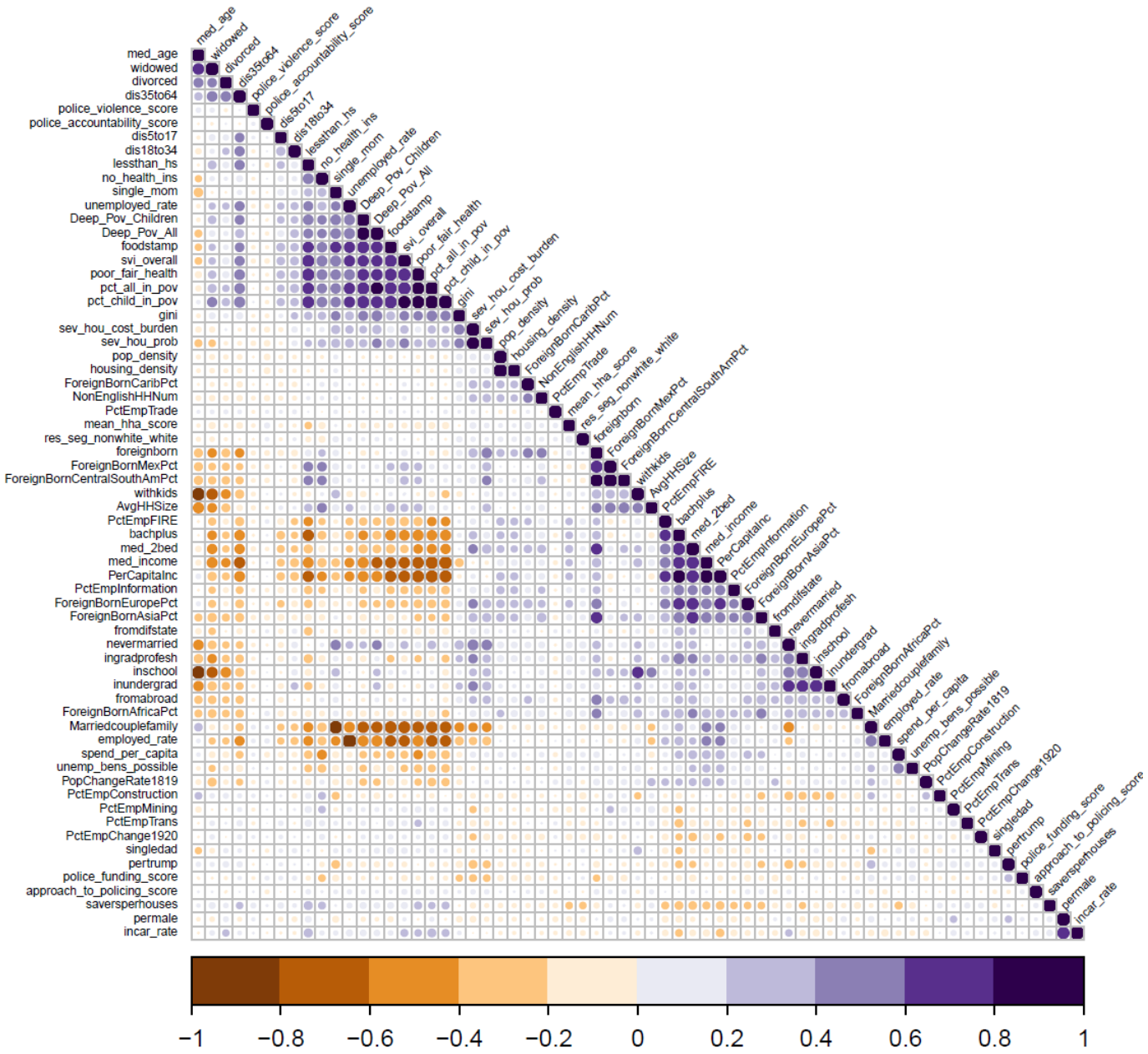

Figure 8: A correlation plot for all features in all categories

This overall correlation plot in Figure **8** allows comparison across different categories of variables. The upper left of the plot shows many significant negative correlations. Notably, poverty-related features are positively correlated with Percent adults reporting poor or fair health, which is a health-related factor. Poverty-related features are also positively correlated with Percent of households qualifying for food stamps, which belongs to the category of Social Safety Net. The lower left part of the plot shows several negative correlations. Specifically, SES features like Median household income, Percent with college or higher education, and employment rate, are negatively correlated with poverty-related variables. Married

couple, which is a basic demographic variable, is also negatively correlated with poverty-related variables. On the one hand, since the observed significant correlations are commonsensical, it implies that our data are quite reliable. On the other hand, the plot shows that some of our features are highly dependent, which would affect our ability to gain insights regarding the most important features contributing to high theft rates.

Scatterplots for Important Features

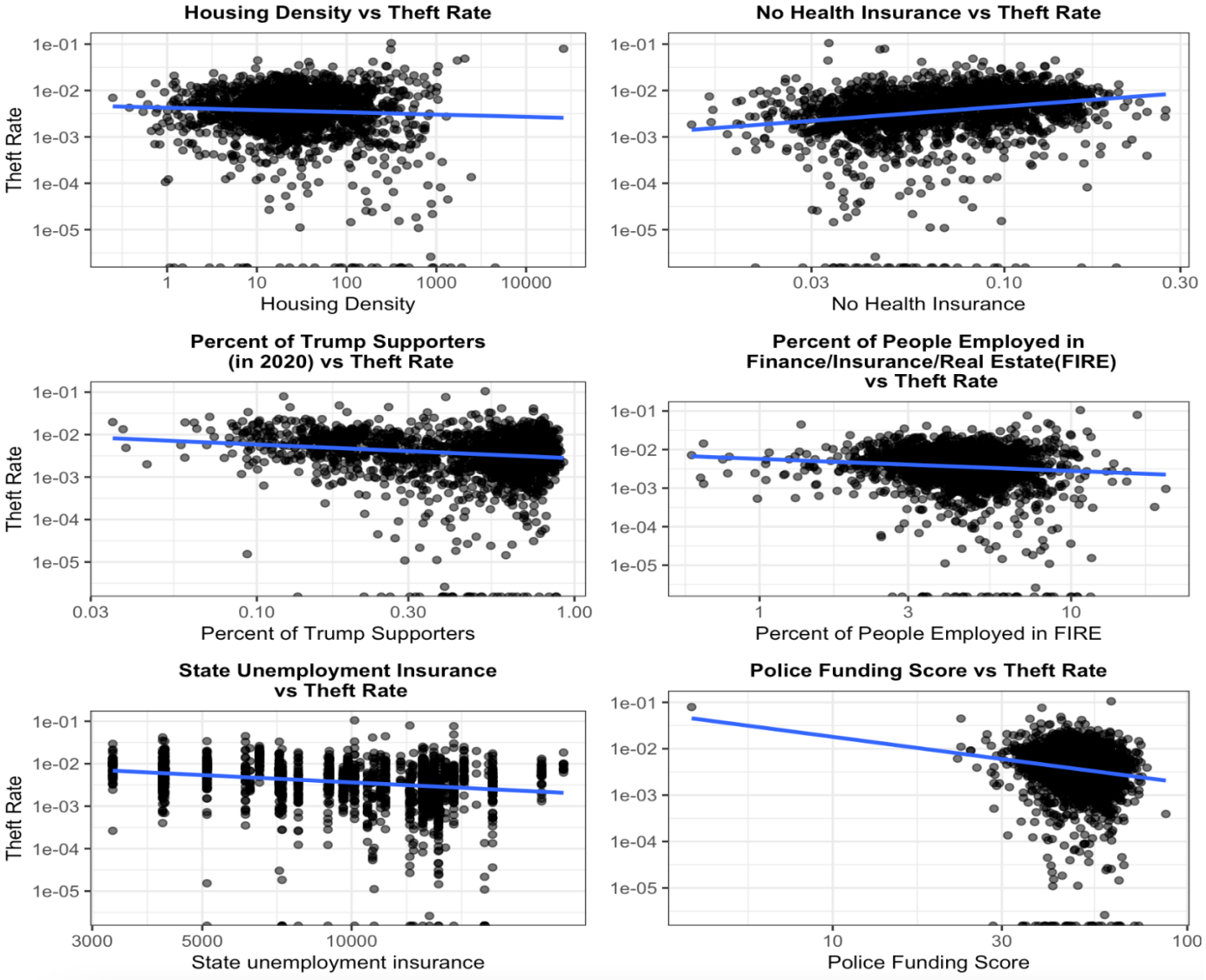

Figure 9: Scatterplots for important features

Figure **9** shows how the theft rate varies with each of the six features, which we selected based on the five models we trained in the next section. The 'No Health Insurance' feature appears to have a relatively strong relationship with the theft rate among the six plots, and it shows a positive correlation. This suggests that the counties with greater percentages of people with no health insurance have higher theft rates. We observed negative correlations between the theft rate and each of the five other features, including Housing density, Percent of Trump supporters, Percent employed in FIRE industries (i.e., finance and insurance, and real estate and rental and leasing), State unemployment insurance, and Police Funding Score.

Modeling

Model Class 1: Regression-Based Methods

To ensure reproducibility, we first set seed. We opted to choose three types of penalized regression models: Ridge, Lasso, and the Elastic Net. Each of these were tuned with cross-validation to select levels of lambda that would most reduce CV error. In the case of the elastic net, an additional tuning parameter, alpha, is used, which could be seen as a way to decide between ridge and lasso's approaches. However, we learned that the data that we collected are not suited for these types of models. For all of these values, the value of Lambda that was chosen was so high that essentially there were no coefficients. They likely did this because the additional variance for including them would overtake any reduction in bias. The models thus seem to find that they could not improve their performance over an intercept-only approach. We provide these results below.

Ridge

Unlike the following two models, Ridge regression does not remove variables from the model: it only penalizes them in order to reduce model complexity, if necessary. This certainly happens in our case, as The coefficients of our variables are so small that they do not truly impact the model. Nevertheless, the variables with the greatest absolute value, in order, are shown in Table **2**.

Feature	Coefficient
housing_density	0
pop_density	0
pertrump	0
unemp_bens_possible	0
police_funding_score	0
gini	0
Marriedcouplefamily	0
no_health_ins	0
pct_child_in_pov	0
svi_overall	0

Table 2: This is a table of the top 10 features selected by Ridge.

You can see on the Ridge Trace plot according to the One Standard Error rule in Figure **10**. The model chosen is the location where these variables are the smallest.

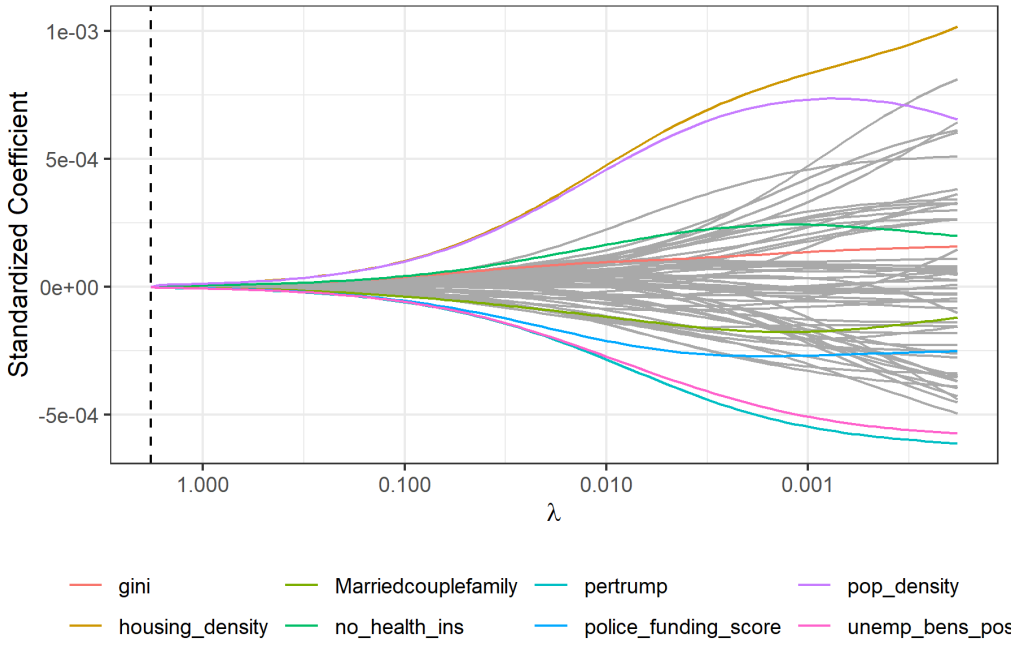

Figure 10: A Ridge trace plot of coefficients, where the dashed line indicates the lambda value chosen using the one-standard-error rule

Like the following two models will show, more complex models using this same methodology would increase the coefficients for these variables. One can determine this simply by looking at the above Trace plot and following the colored lines to the right. Thus, this trace plot could be used to direct future research on a potential positive relationship of housing and population density, as well as a lack of health insurance, on theft. Similarly, theft could have a negative relationship with Trump support, the strength of unemployment benefits, and police funding. Nevertheless, in this case, cross-validated ridge regression found that these later models did not actually improve CV error and were thus excluded from the final model.

Lasso

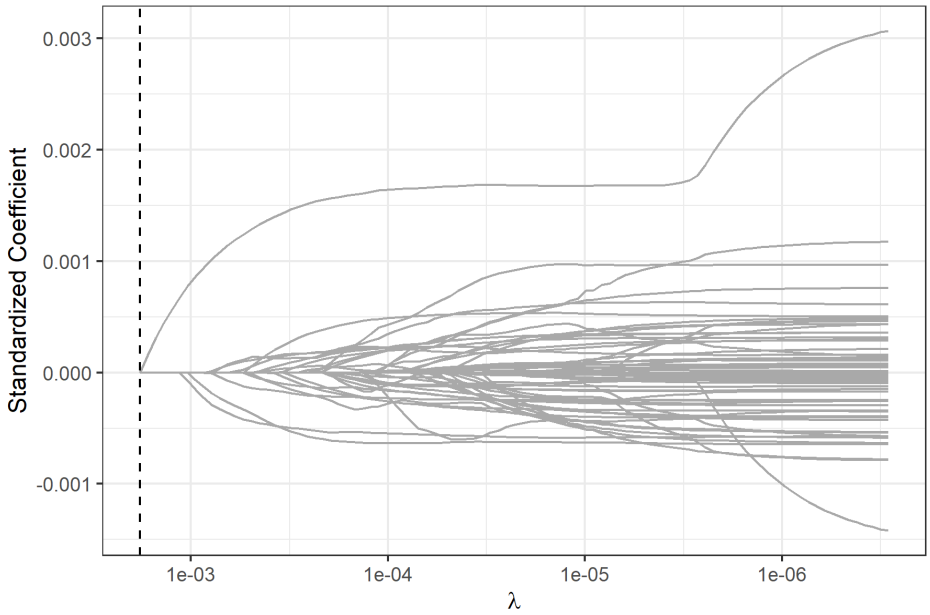

Figure 11: A Lasso trace plot of coefficients, where the dashed line indicates the lambda value chosen using the one-standard-error rule

The gray plot in Figure **11** shows that, according to the one-standard-error rule, lasso regression did not select any variables to include. This means that the value of lambda chosen was so high that the sparse methodology that lasso uses found that no variables could help improve CV error. Specifically, the increasing variance caused by increasing the model complexity (by including the variables) would be greater than any decrease in bias.

Elastic Net

Lastly, we used the elastic net model to see if anything might change by creating a custom model somewhere between Ridge and lasso on the value of alpha. The result of this calculation is as follows:

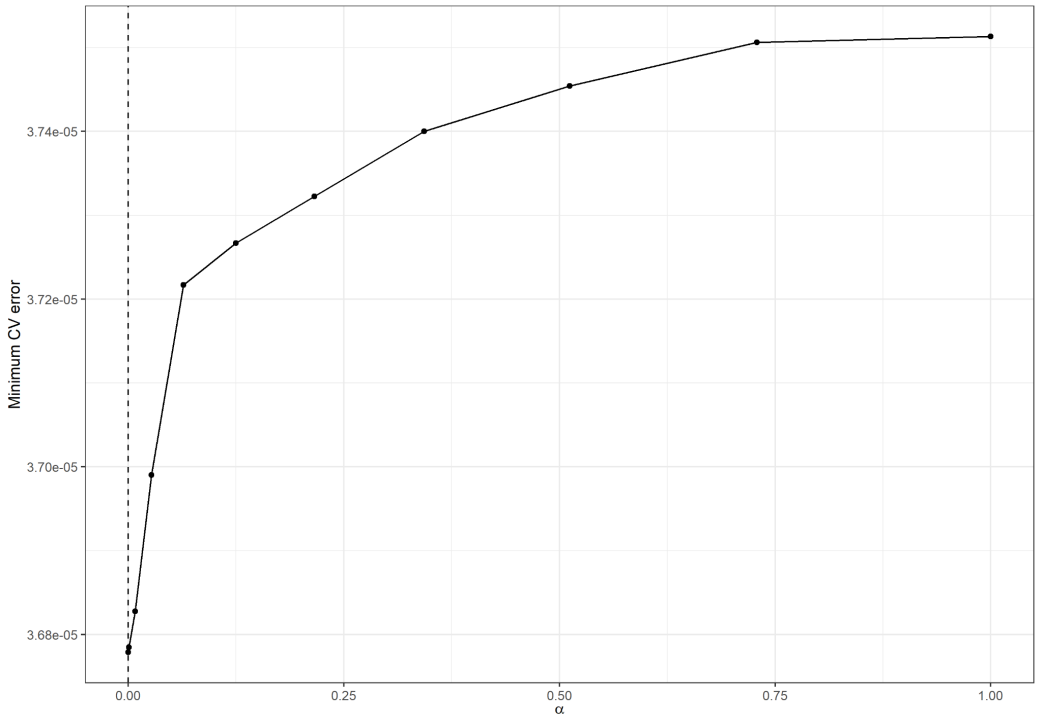

Figure 12: A plot of the minimum CV error for each value of alpha

As we can see from the plot in Figure **12**, that the model was heavily favored towards Ridge, and against a sparse model like seen in lasso. This can be found by seeing that the lowest CV error was found with an alpha level of 0, which is equivalent to ridge. Thus, the elastic net model behaves very much like the ridge model. The resulting Trace plot can be seen below.

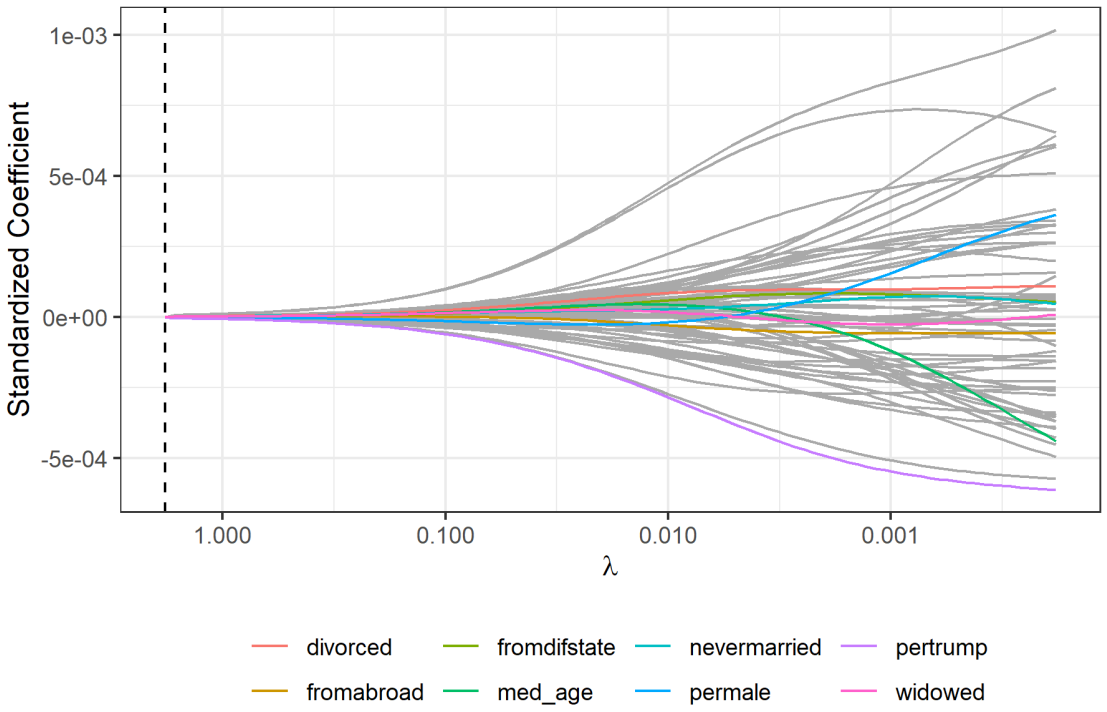

Figure 13: A Elastic Net trace plot of coefficients, where the dashed line indicates the lambda value chosen using the one-standard-error rule

Nevertheless, in Figure **13**, we again see that the one standard error rule has chosen a level of Lambda that is so great that the penalty on are coefficients essentially reduces them to effectively zero. This once again confirms our finding that the variance in our data is so high that increasing the bias to reduce it is of no use when it comes to reducing errors. Interestingly, more complex versions of this model find that the proportion of males and divorcees in a county is associated with a higher rate of theft, and Trump support, along with increase in age, is associated with a lower one. Again, these are not conclusions: instead they should only be used for future research questions.

Model Class 2: Tree-Based Methods

It appears that the penalized regression models above were unlikely able to outperform the intercept-only model, given the fact that the coefficients were chosen to be so greatly penalized (or excluded altogether). Naturally, we felt the need to turn to tree-based methods to see if this will help us understand the underlying true model behind theft.

Random Forest

We first set a seed to make the results reproducible. We tuned the random forest model by finding the optimal number of features to consider at each tree split. This was accomplished by training the model on different values of m, ranging from 6 to 20. Instead of using a range from 1 to 65, we chose the range based on previous efforts on tuning the model, which suggested that a relatively small number of features with increments of size one was the most efficient way to go. According to Figure **14**, which shows the OOB error as a function of the number of trees, the error stabilizes and states flat when B is large enough

(around 100 in this case). However, since we did not have to save computational resources, we chose the default number of B, which is 500, to train our model.

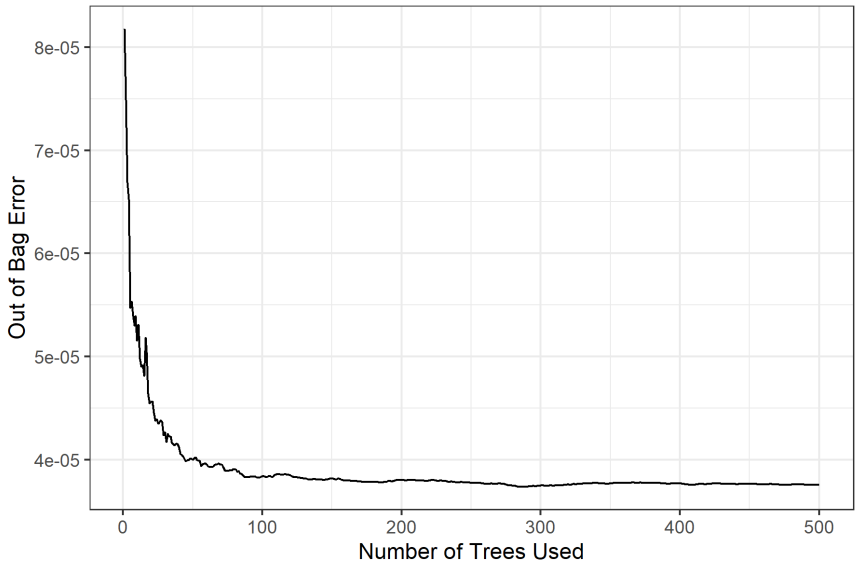

Figure 14: A visualization that shows the OOB error as a function of the number of trees.

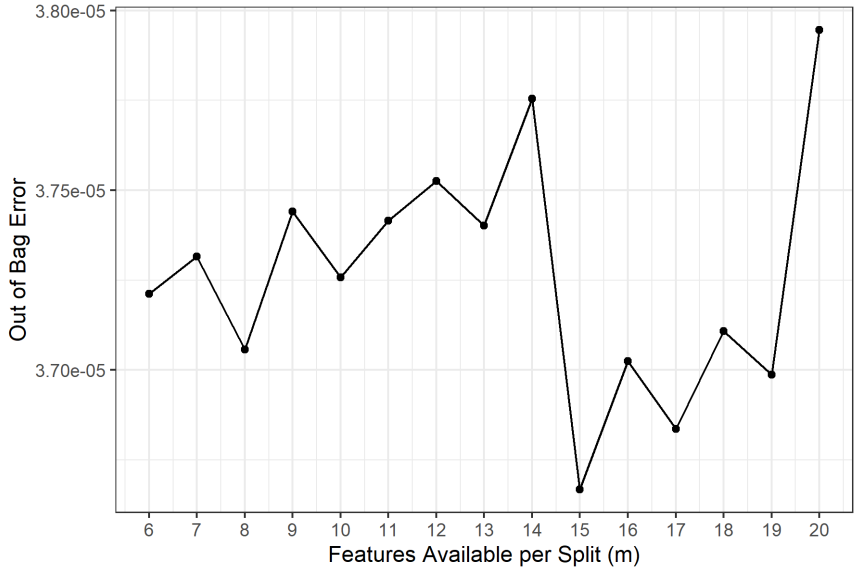

Figure 15: A plot that illustrates the out-of-bag error for each value of m.

The out-of-bag error for each value of m can be observed in Figure **15**, which shows that the out-of-bag error is minimized at the value of m = 15. We then fit our tree using the m value specified above and assessed variable importance based on the OOB variable importance. Looking at the error based chart on the left in Figure **16**, we see that the most important variables are, in order, Possible Unemployment Benefits, Trump support, Government spending per capita, Percent of people with a disability from ages 35 to 64, police accountability, percent with at least a bachelor's degree, percent of children in poverty, population density, and the median cost to rent a two-bedroom apartment. The Purity based chart on the right is included for comparison.

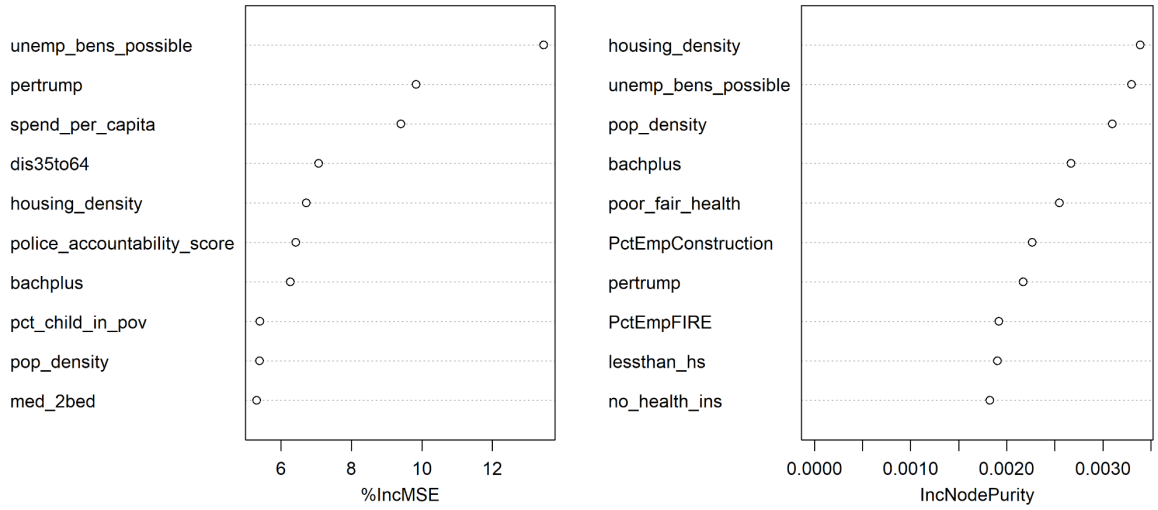

Figure 16: A plot based on two feature importance measures, which are OOB-based importance and Purity based Importance

Boosting

Finally, we used a Gradient Boosting model to approach the problem from another angle by having shallower tree depths–and to allow for specific variable importance plots, which are produced below. First, we trained three different models based on depths 1through thre3, with depth 3 providing the lowest error, found at 21 trees. This was determined using the plot in Figure **16**, which compares depth levels 1 through 3 against each other over the course of 1 through 1000 trees.

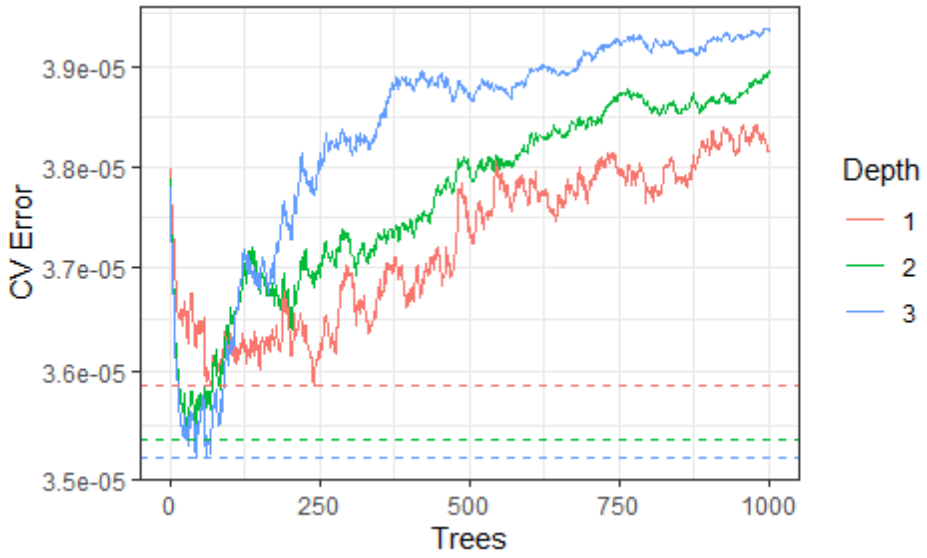

Figure 16: A plot of CV errors for each of the three interaction depths.

This gradient boosting model identified the following variables as having the greatest relative influence. The top six of these and their individual relative influence plots are given below in Table **3** and Figure **17**, respectively.

Variable	Relative influence
bachplus	22.800
PctEmpFIRE	9.707
unemp_bens_possible	9.269
housing_density	4.708
pertrump	4.603
poor_fair_health	3.635
pop_density	3.632
police_funding_score	3.370
withkids	3.279
dis35to64	3.241
lessthan_hs	3.088
no_health_ins	2.566

Table 3: This is a table of the top 12 features according to Boosting..

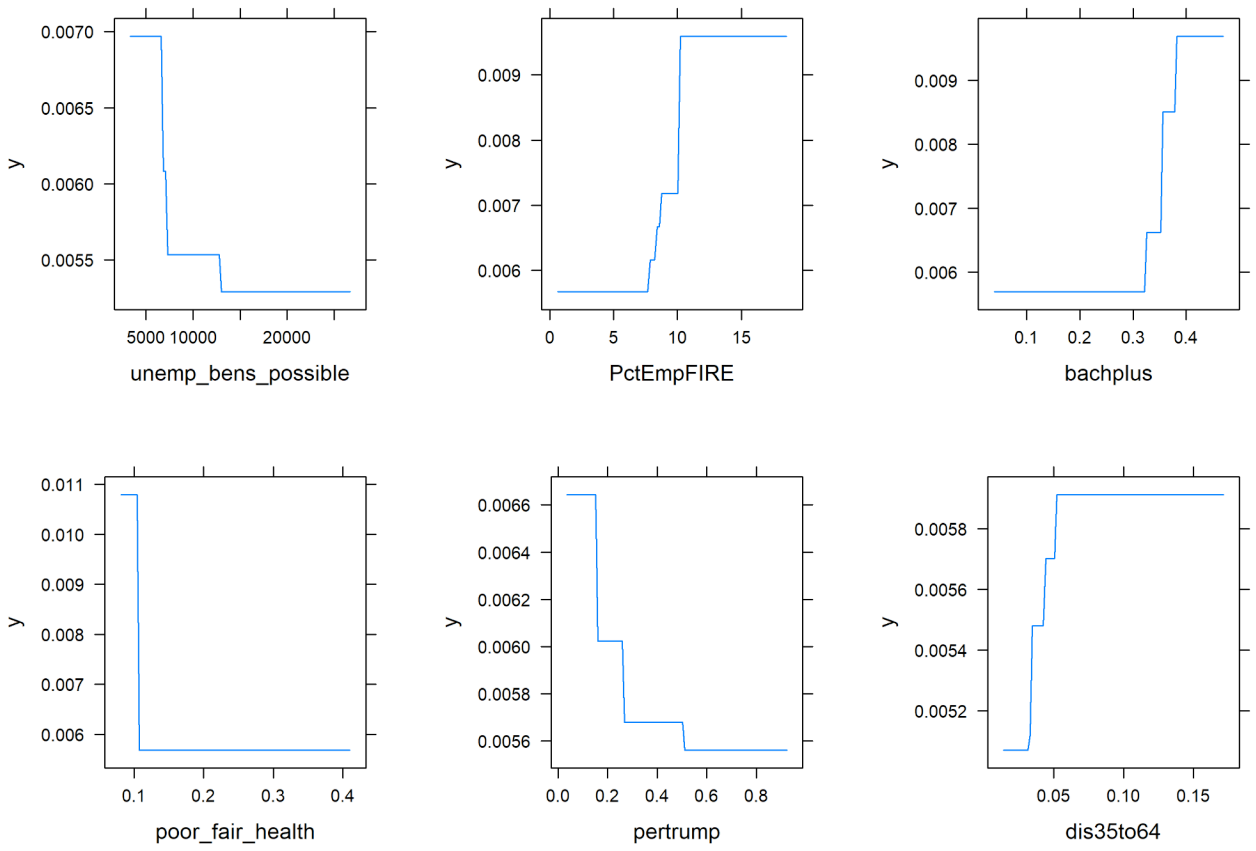

Figure 17: Partial dependence plots for the top six features

These relative influence plots function in a sense like correlation plots, with each break in the line associated with a tree-split on that feature. Some of these relationships are surprising, others are not. For instance, it appears that the estimated theft rate decreases as more unemployment benefits are made available, as has been seen in other models above. The second plot captures the percentage of people in the county in certain professional jobs, such as **F**inance and **I**nsurance, and **R**eal **E**state and rental and leasing. In addition, the rate of people with at least a bachelor's degree seems associated with a higher

theft rate as well. The percent of people who report fair or poor health is associated with a lower theft rate, as is support for president Trump, which too is a recurring factor. Lastly, the percentage of medium aged adults, 35 through 64, with a disability, is associated with a higher theft rate.

Conclusions

Method comparison

After we ran all the models, we compared them against an intercept-only model. The results can be found in Table **4**. We were tipped off that many of our models might not be any better than this one through the fact that the penalized regression models derived a penalty that was so great as to render the formula to have little to no weight to the coefficients.

Model	Test_RMSE
Boosting	0.0047690
Intercept-only	0.0048906
Ridge	0.0049133
Lasso	0.0049133
Elastic_Net	0.0049133
Random_Forest	0.0049465

Table 4: This is a table of each model's root mean square error for the testing data.

Because it's results are on the same scale as the response variable, we chose to evaluate our models using the root mean squared (RMSE) error calculation. The mean of the response value in the training data is 0.005757227. Our models did not perform better than the intercept-only version, except for the gradient boosting model, which did so only slightly. We can see that for ridge, lasso and elastic net, they all have the same RMSE, because they derived essentially the same intercept-only model through penalizing the coefficients down to effectively 0.

In this case, as is often seen, gradient boosting provides the lowest error. Gradient boosting makes predictions by utilizing weak learners that are successive, and get closer to the goal by reducing residuals on each step. It appears the random forest method fell victim to what the penalized regression based methods were attempting to avoid. By reducing the variance, the bias may have increased by a comparatively higher magnitude, rendering the model moot. However, it is possible that the boosting method was able to reduce both bias and variance to marginally improve over the intercept-only.

Takeaways

The challenge of finding ways to reduce theft in particular, and crime in general, is naturally a very big one. However, we were surprised to see just how challenging it is to see trends in the data. Only one of our models was able to improve over The intercept-only model. Admittedly, we were quite surprised by this reality, as we included many variables which we believed could theoretically be involved with reducing (or increasing) theft. We take our results as essentially null results, and treat this almost as a

failed replication. What does this mean if the measures we collected do not appear to have a strong relationship with, or cannot be aggregated into a successful predictive model to understand changes in theft rates between areas? It likely means several things:

1. There are greater complexities under the rate of theft that we were not able to capture.
2. Better data to drill down to smaller jurisdictions or areas, instead of counties, may be needed to make a proper analysis. We suspect that with many more observations, these models would be able to create an algorithm strong enough to beat the intercept-only.
3. Common assumptions about what drives the theft rate may not be correct. What our chosen models have told us is that simply knowing the features that we provided on a county level was not enough for the models we chose to make good predictions on the test data.

Limitations

Perhaps the greatest limitation to our analysis was the fact that our algorithms do not produce predictions better than simply using the intercept-only model, with the exception of gradient boosting, which is only marginally more accurate. Thus, any data we derive in this analysis should be seen as only hints and requests for future research. In addition, we did the analysis on the county level. We imagine that predictive models would have had a better performance if we were able to analyze smaller communities and get more specific data. A further limitation was that many of our features had incomplete data. This stems from a data-collection problem that many state and local governments face. A further limitation is the fact that we were using data from different years to predict the theft rate in 2020. This may have contributed to bias.

Follow-ups

We recommend that the next analysis on this topic be on a much more local level. The first step for this kind of approach could be to find which cities have the best data collection and reporting. Perhaps it is best to create a model from a few of these cities and then use that to predict crime in another city with similarly good data collection practices. In addition, while we were able to recover some lost counties by using a random forest method to impute possible missing values in the incarceration rate, we were still missing others. It is unclear if these additional counties would improve the models. Nevertheless, future analyses must be very careful when including data which might reduce the number of viable observations. If possible, it is recommended that a few large data sets with many features and observations be used for ease of analysis and perhaps more consistency in the data.

Appendix

Explanatory Variables

Below are the 65 features we used for analysis. Words written in parentheses represent feature names used in R analysis. All features are continuous.

Basic Demographics

- **Median age (med_age):** Median age of county
- **Gender (permale):** Percentage of residents who are male
- **Percent divorced (divorced):** Percentage of residents who are divorced
- **Percent widowed (widowed):** Percentage of residents who are widowed
- **Percent never married (nevermarried):** Percentage of residents who are never married
- **Percent of Trump supporters (pertrump):** Percent voting for Trump in 2020
- **Population density (pop_density):** Total population divided by square miles
- **Housing density[3] (housing_density)**: Number of housing units divided by square miles
- **Residential segregation (res_seg_nonwhite_white)**: Index of dissimilarity where higher values indicate greater residential segregation between White and non-White county residents.
- **Married couple (Marriedcouplefamily)**: Percent of households that contain a married couple living together
- **Percent of persons born in Europe (ForeignBornEuropePct)**: Average percent of persons born in Europe from 2015 to 2019
- **Percent of persons born in Mexico (ForeignBornMexPct)**: Average percent of persons born in Mexico from 2015 to 2019
- **Percent of persons born in the Caribbean (ForeignBornCaribPct)**: Average percent of persons born in Caribbean from 2015 to 2019
- **Percent of persons born in Central or South America (ForeignBornCentralSouthAmPct)**: Average percent of persons born in Central or South America from 2015 to 2019
- **Percent of persons born in Asia (ForeignBornAsiaPct)**: Average percent of persons born in Asia from 2015 to 2019
- **Percent of persons born in Africa (ForeignBornAfricaPct)**: Average percent of persons born in Africa from 2015 to 2019
- **Number of non-English speaking households (NonEnglishHHNum)**: Average number of non-English speaking households from 2015 to 2019
- **Average household size (AvgHHSize)**: Average household size from 2015 to 2019
- **Population change rate (PopChangeRate1819)**: Population change rate from 2018 to 2019
- **Percent of household with own children (withkids)**: Percent of households that contain children of the householder
- **Percent of households with single mothers (single_mom)**: Percent of households with single mothers
- **Percent of households with single fathers (singledad)**: Percent of households with single fathers
- **Percent of foreign-borns (foreignborn)**: Percent of persons born outside the United States

[3] This is not official definition of housing density.

- **Percent of people moved from other states (fromdifstate)**: Percent of people who moved from a different state in the last year
- **Percent of people moved from abroad (fromabroad)**: Percent of people who moved from a different country in the last year

Socioeconomic Status (SES)

- **Percent with less than high school education (lessthan_hs)**: Percent of people with less than high school education
- **Percent with college or higher education (bachplus)**: Percent of people with a Bachelor's degree or higher
- **Unemployment rate (unemployed_rate)**: Annual rate of unemployment calculated by the number of unemployed residents in a county divided by the total number of workers available in labor force
- **Employment rate (employed_rate)**: Annual rate of employment calculated by the number of employed residents in a county divided by the total number of workers available in labor force
- **Cost of living (med_2bed)**: Median monthly rent for a two-bedroom apartment
- **Gini index (gini)**: a summary measure of income inequality that ranges from 0, indicating perfect equality (where everyone receives an equal share), to 1, perfect inequality (where only one recipient or group of recipients receives all the income).
- **Social Vulnerability Index (svi_overall)**: a measure of social need in a given county for emergency response that involve 4 major themes: socioeconomic status, household composition and disability, race/ethnicity and language, and housing or transportation status.
- **Percent of people in poverty (pct_all_in_pov)**: Percent of people living in families with total income below the official poverty threshold.
- **Percent of children in poverty (pct_child_in_pov)**: Percent of children under 18 years old in families with incomes below 100% of the federal poverty level
- **Median household income (med_income)**: Median household income in the past 12 months
- **Percent of households with severe housing cost burden (sev_hou_cost_burden)**: Percent of households that spend 50% or more of their household income on housing
- **Percent of households with severe housing problems (sev_hou_prob)**: Percent of households with at least 1 of 4 housing problems: overcrowding, high housing costs, lack of kitchen facilities, or lack of plumbing facilities
- **Percent of employment change (PctEmpChange1920)**: Percent of employment change from 2019 to 2020
- **Percent employed in construction (PctEmpConstruction)**: Average percent of the civilian labor force (aged 16 and over) employed in construction from 2015 to 2019
- **Percent employed in mining, quarrying, oil and gas extraction (PctEmpMining):** Average percent of the civilian labor force (aged 16 and over) employed in mining, quarrying, oil and gas extraction from 2015 to 2019
- **Percent employed in wholesale and retail trade (PctEmpTrade)**: Average percent of the civilian labor force (aged 16 and over) employed in wholesale and retail trade from 2015 to 2019
- **Percent of employed in transportation, warehousing and utilities (PctEmpTrans)**: Average percent of the civilian labor force (aged 16 and over) employed in transportation, warehousing and utilities from 2015 to 2019

- **Percent employed in information services (PctEmpInformation)**: Average percent of the civilian labor force (aged 16 and over) employed in information services from 2015 to 2019
- **Percent employed in finance and insurance, and real estate and rental and leasing (PctEmpFIRE)**: Average percent of the civilian labor force (aged 16 and over) employed in finance and insurance, and real estate and rental and leasing from 2015 to 2019
- **Per capita income in the past 12 months (PerCapitaInc)**: Average per capita income in the past 12 months (in 2019 inflation adjusted dollars) from 2015 to 2019
- **Percent of people in deep poverty (Deep_Pov_All)**: Average percent of residents in deep poverty from 2015 to 2019
- **Percent of children in deep poverty (Deep_Pov_Children)**: Average percent of children in deep poverty from 2015 to 2019
- **Percent of people enrolled in school (inschool)**: Percent of population aged 3 and over enrolled in school
- **Percent of people enrolled graduate or professional schools (ingradprofesh)**: Percent of population enrolled in Graduate or Professional School
- **Percent of people enrolled in college (inundergrad)**: Percent of population enrolled in an Undergraduate program

Social Safety Net

- **State unemployment insurance (unemp_bens_possible)**: Average total amount of benefits in dollars one can receive from unemployment insurance by state (Amount of weeks multiplied by amount to receive per week.)
- **State and local government spending on people (spend_per_capita)**: the annual per-capita amount spent on schools, health care services, and general administration (among other activities in the general government sector) but exclude government-run liquor stores, utilities, and insurance trusts, which are accounted for separately in the census.
- **Percent of people qualifying for Saver's Credit (saversperhouses)**: Percent of people who qualify for the Saver's Credit (tax break for saving for retirement within relatively low-income limits).
- **No health insurance (no_health_ins)**: Percentage of people with no health insurance.
- **Percent of households qualifying for food stamps (foodstamp)**: Percent of households qualifying for Supplemental Security Income (SSI), Cash Public Assistance Income, or Food Stamps/Snap in the Past 12 Months

Criminal Justice Response

- **Incarceration rate (incar_rate)**: Percentage of adults in correctional facilities in 2020.
- **Police Violence Score[4] (police_violence_score)**: Percentage calculated by taking the simple average across: Percentile Less Lethal Force Used per Arrest, Percentile Deadly Force Used per Arrest, Percentile Unarmed Civilians Killed or Seriously Injured, and Percentile Racial Disparities in Arrests and Deadly Force*
- **Police Accountability Score (police_accountability_score)**: Percentage calculated by taking differently weighted average across the following: Percentile Civilian Complaints Sustained (50%

[4] For complete definitions of variables mentioned here, please refer to the Police Scorecard Project.

weight) and the average (50% weight) across Percent Discrimination complaints sustained, Excessive Force Complaints Sustained, and Percent Criminal Complaints Sustained.

- **Approach to Policing (approach_to_policing_score)**: Percentage calculated by taking the simple average across the following: Percentile Low Level Arrests per Population and Percent Homicides Cleared
- **Police Funding Score (police_funding_score)**: Percentage calculated by taking the simple average across the following: Percentile Police Funding per Population, Percentile Number of Officers per Population, Percentile Average Funds Spent on Misconduct Settlements, Percentile Funds Taken From Communities in Fines, and Forfeitures per Population

Health-related Factors

- **Hardest Hit Area (COVID) score (mean_hha_score)**: Areas designated as either a "sustained hotspot," or a "hotspot," on the COVID-19 Community Profile Report, Area of Concern Continuum by County dataset provided by the U.S. Department of Health and Human Services (HHS). A "sustained hotspot" is defined by HHS as a community that has "a high sustained case burden and may be higher risk for experiencing health care limitations." Hotspots are defined by HHS as "communities that have reached a threshold of disease activity considered as being of high burden."
- **Percent adults reporting poor or fair health (poor_fair_health)**: Percentage of adults reporting fair or poor health (age-adjusted)
- **Percent of 5-17yo people with disability (dis5to17)**: Percent of residents aged between 5 and 17 having serious difficulty with any one of four basic areas of functioning – hearing, vision, cognition, and ambulation.
- **Percent of 18-34yo people with disability (dis18to34)**: Percent of residents aged between 18 and 34 having serious difficulty with any one of four basic areas of functioning – hearing, vision, cognition, and ambulation.
- **Percent of 35-64yo people with disability (dis35to64)**: Percent of residents aged between 34 and 64 having serious difficulty with any one of four basic areas of functioning – hearing, vision, cognition, and ambulation.