CRIME DISSIPATION USING CRIME DATA

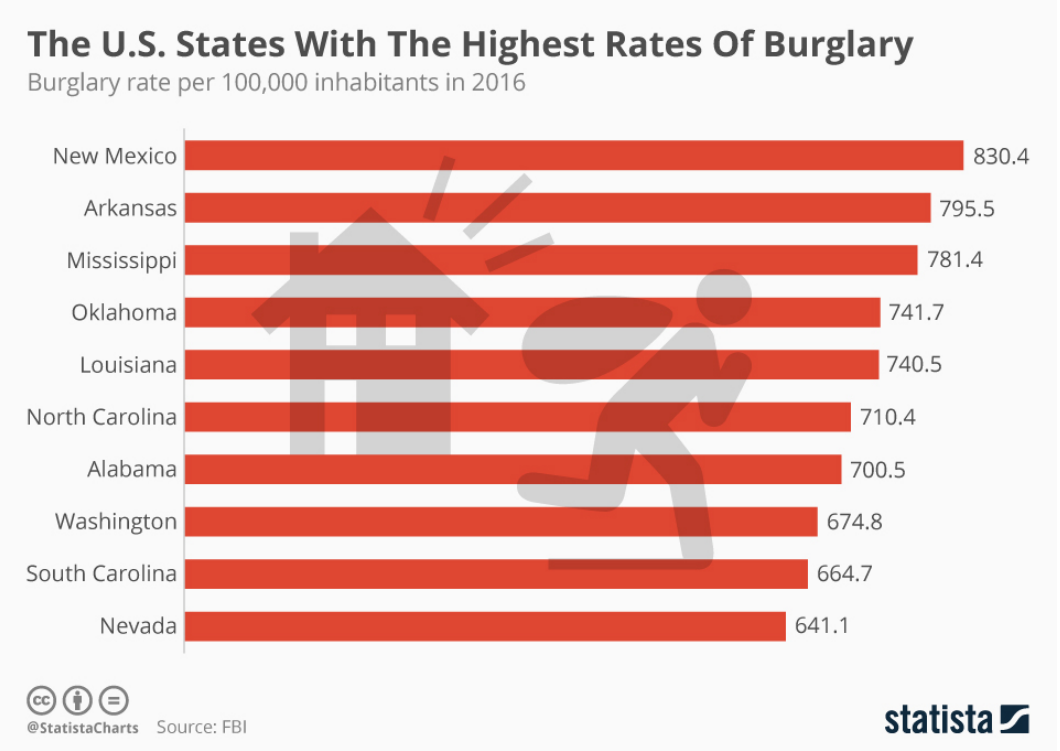
ANNA DAMIR

OVERVIEW

- Explanation of The Problem
- The Solution
- How to use the model(s)
- Why these models
- The Dataset
- What to do with the model results/looking towards the future

THE ISSUE

- While the crime rate has decreased over the last few decades, people still do not feel safe
- 22% of Americans in 2017 said their household was victimized by crime in the past year
- Nonprofits have been the most effective in reducing crime on the local level by providing programs to victims of families, rehabilitation of and education for inmates, and programs for at-risk youth
- These programs cost money, and we want to know the cities to focus on: those with the highest crime rates, should be higher on the priority list than those where crime is less likely to be committed.



https://www.statista.com/chart/12959/the-us-states-with-the-highest-rates-of-burglary/

THE SOLUTION

- Goal is to predict crime data for cities far before the actual crime data is acquired, which due to red tape takes half a year or more
- This will allow for mobilization of money and people far earlier, so that programs can start making an impact immediately
- Comparison of 3 different models I created and the recommendation of which to use for what

THE SOLUTION SPECIFICS

3 Models Tested

Linear Regression

- Simplest, most general regression
- Least accurate, but fast results
- Makes too many assumptions that aren't good for this dataset

Ridge Regression

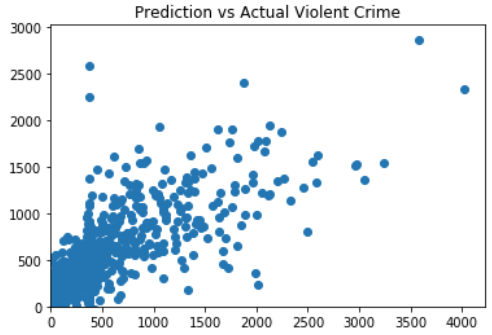
- Deals with correlation within the dataset
  - Will lead to more accurate results when predicting than linear alone

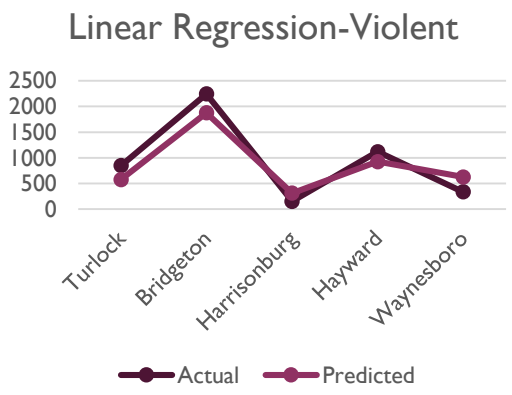
Random Forest Regression

- Most comprehensive regression
- Shows non-linear correlation between variables, if any
- Does take longer to run due to nature of model

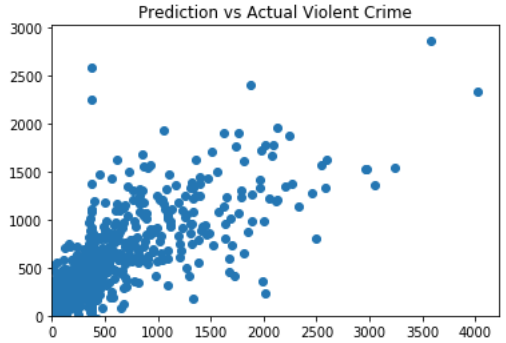
THE SOLUTION SPECIFICS CONT.

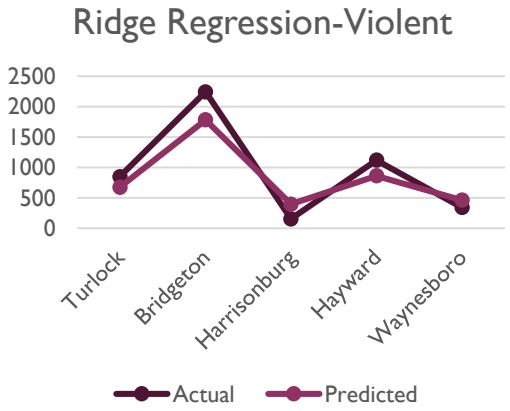
Linear Regression



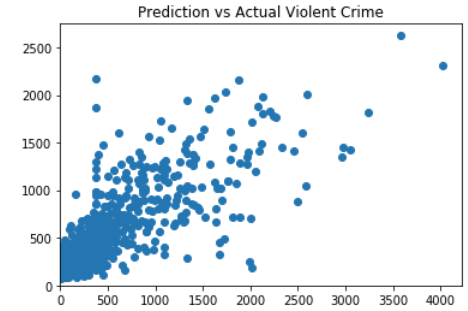


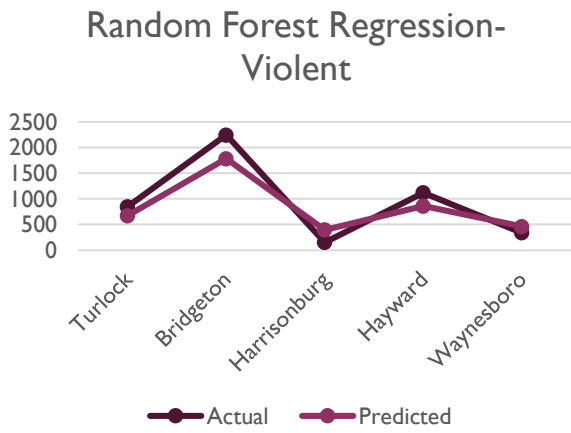
Ridge Regression





Random Forest Regression





THE SOLUTION SPECIFICS CONT.

Linear Regression

- Violent Crime

```
Root Mean Squared Error: 392.1659231096209
```

	Cities	Actual	Predicted
0	Turlock	847.77	577.00
1	Bridgeton	2241.85	1875.08
2	Harrisonburg	151.78	311.92
3	Hayward	1118.18	925.06
4	Waynesboro	337.96	627.34

- Nonviolent Crime

```
Root Mean Squared Error: 2139.8709139461266
```

	Cities	Actual	Predicted
0	Turlock	7632.05	5475.45
1	Bridgeton	8005.85	8507.37
2	Harrisonburg	4264.75	5080.85
3	Hayward	6251.13	6193.29
4	Waynesboro	4198.13	5909.05

Ridge Regression

- Violent Crime

```
Root Mean Squared Error: 391.75739545423096
```

	Cities	Actual	Predicted
0	Turlock	847.77	571.65
1	Bridgeton	2241.85	1863.30
2	Harrisonburg	151.78	303.16
3	Hayward	1118.18	916.16
4	Waynesboro	337.96	616.25

- Nonviolent Crime

```
Root Mean Squared Error: 2138.087273996554
```

	Cities	Actual	Predicted
0	Turlock	7632.05	5483.65
1	Bridgeton	8005.85	8451.01
2	Harrisonburg	4264.75	4926.89
3	Hayward	6251.13	6140.77
4	Waynesboro	4198.13	5827.57

Random Forest Regression

- Violent Crime

```
Root Mean Squared Error: 370.27503703261243
```

	Cities	Actual	Predicted
0	Turlock	847.77	673.53
1	Bridgeton	2241.85	1781.98
2	Harrisonburg	151.78	394.59
3	Hayward	1118.18	859.89
4	Waynesboro	337.96	460.66

- Nonviolent Crime

```
Root Mean Squared Error: 2222.28463409814
```

	Cities	Actual	Predicted
0	Turlock	7632.05	5513.71
1	Bridgeton	8005.85	7668.40
2	Harrisonburg	4264.75	4669.15
3	Hayward	6251.13	6301.94
4	Waynesboro	4198.13	5350.08

THE SOLUTION SPECIFICS CONT.

- For violent crime, the best model to use is random forest regression
  - Highest r^2 value and lowest Root Mean Squared Error
  - Using this on data will not give predictions true enough to determine how much crime occurred but is accurate enough to use as guidelines for picking which cities need the most help first
- For non-violent crime, I do not recommend using any of the models to predict
  - Far too much error in all of the predictions
  - The variables we tested for in data set are not good indicators of nonviolent crime
  - However, if model must be used, ridge regression model is the best, though answers will not be as accurate as violent crime

HOW TO USE THESE MODELS

- What I've provided:
  - The models, already trained with older data, with the assumption that while socioeconomic data may change, what contributes to crime does not change through the years
- What you should provide:
  - The socioeconomic data for the place(s) you are trying to predict (i.e. city population, breakdown of families and types, general census type data)
    - The more complete and recent the data is the better
- You enter the up to date data into the model, and the prediction results are provided
  - Organized top 20 cities by highest predicted crime rate

WHY THESE MODELS

- Wanted to find out the best correlations possible between census level data and crime rates in communities
- Begun with correlating individual variables to the ones to be predicted, the crime rates
- Start out with the most simple prediction model, despite it not being the most accurate, and repeat with better possible models for the given data, chose ridge and random forest which tested for slightly different things in the data, ridge for relationships of independent variables, and random forest for nonlinear relationships
- Choose the best of them possible. Accuracy is not as high as wanted due to difficulty of predicting social science type actions. By nature, humans are less predictable than machine. Also issues with dataset….

	R-values
pctKids2Par	-0.684059
pct2Par	-0.649762
pctWhite	-0.647164
pct12-17w2Par	-0.616963
pctKids-4w2Par	-0.611884
pctWdiv	-0.534563
pctPersOwnOccup	-0.494272
pctHousOwnerOccup	-0.443022
medFamIncome	-0.394124
medIncome	-0.380866
rapesPerPop	0.583146
pctBlack	0.590580
autoTheftPerPop	0.600200
nonViolPerPop	0.627124
murdPerPop	0.635940
burglPerPop	0.676466
pctKidsBornNevrMarr	0.686665

THE DATA SET I USED

- Used 1990 US Census, 1990 Law Enforcement Management and Admin Stats Survey, 1995 US FBI Uniform Crime Report
- Older data, but though the data is old, we assume the model is still fine since the causes behind the crime stay the same though the socioeconomic data may differ, but by incorporating some level of new data into the model, we adjust for this, and provide predictions based on the newer data
- Old data had lots of missing data, no easy way to fill in blanks, better dataset would've created a higher accuracy model
- Missing data also led to assumptions being made prior to models being run, led to less accurate models, but still within reason for prediction.

HOW TO USE MODEL RESULTS & LOOKING TO THE FUTURE

- The model will give predicted crime values for the cities entered within
  - The cities with highest crime rate are output by model and would be those in most need
    - Mobilize those cities first. Focus on funding them before others and getting the community to understand what is needed.
  - Population is also included because while the crime number may be large, you may be wanting to look for the highest ratio of crime to population. Large cities inherently will have more crime by the number, but you cannot let small cities with high rates run under the radar.
- Use the single correlation of variables to get more complete data in the future
- If more complete and recent dataset is found, it's easy to retrain the model. Recommendation of using Random Forest would stay the same.
- Variables beyond the given dataset would be interesting to explore: how many visitors a certain city has, if data is available, narrow down to certain times of year, or even certain neighborhoods in cities