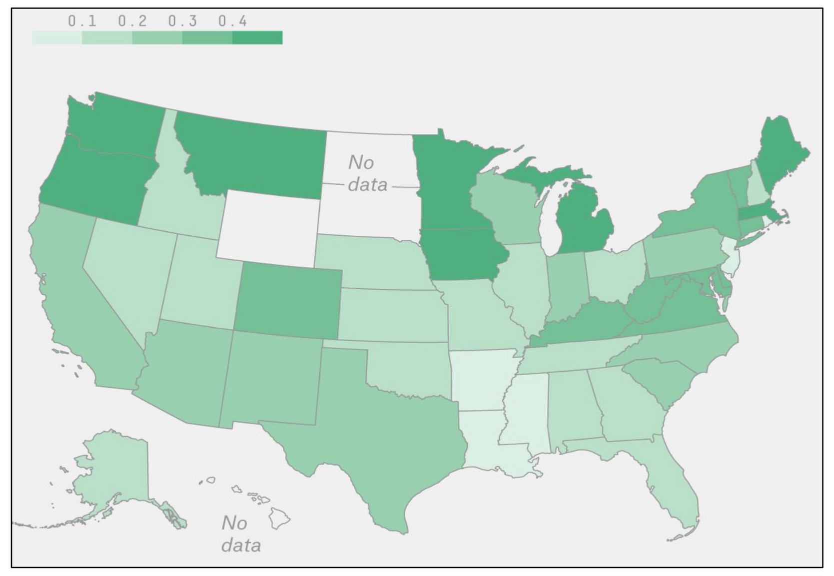
Predicting Hate Crime Rates Following the 2016 Election

Taylor Ayers, Jessy Ha, Botagoz Tusmagambet, Gianna White, Zhaoqi Wu

Introduction

Data collected by the FBI and the Southern Poverty Law Center indicate increased hate crime incidents across the United States following the 2016 presidential election (Figure 1). Hate crime rates vary from state to state. However, it is unclear why some states experienced more incidents of hate crime than other states.



https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/

Figure 1.Post-election hate incident rates. Hate incidents per 100,000 residents 10 days after Nov. 8,2016 [1].

Can we predict hate crime rates per state after the 2016 election? We aim to construct a model that determines the association between social and economic measurements in each state and its rate of crime.

Data

Source of compiled hate crime data [2]: FBI, 2010-2015 and Southern Poverty Law Center (SPLC), Nov. 9-18, 2016

Note: data are not directly comparable. The FBI only receives reports of hate crimes, while the SPLC also records hate incidents, "which may or may not be severe enough to be prosecuted as a crime."

Table 1: Predictor Variables

Predictor Variables	Description
Median Household Income	Median reported household income for 2016
Share Unemployed Seasonal	Percentage of population unemployed September 2016
Share population in Metro Areas	Percentage of population living in metropolitan areas in 2015
Share Population with High School Degrees	Share of adults 25 and older with a high-school degree, 2009
Share Non-citizen	Percentage of non-citizen population in 2015
Share White Poverty	Percent of white residents living in poverty in 2015
GINI Index	Income inequality predictor
Share Non-white	Percentage of non-white population in 2015
Share Voters Voted Trump	Percentage of 2016 voted who voted for Trump
Avg Hate Crimes per 100k FBI	Average annual hate crimes reported to the FBI per 100,000 population 2010-2015

We applied a log-transformation to the predictor variables to identify outliers:

- 'Share Voters Voted Trump' < 0.2
- 'Gini Index' > 0.5

Table 2: Response Variable

Response Variable	Description
Hate Crimes per 100k SPLC	Hate crimes per 100,000 population, collected Nov. 9-18, 2016 by SPLC

The following states were excluded from the analysis due to missing data: Hawaii, North Dakota, South Dakota, Wyoming, Maine and Mississippi.

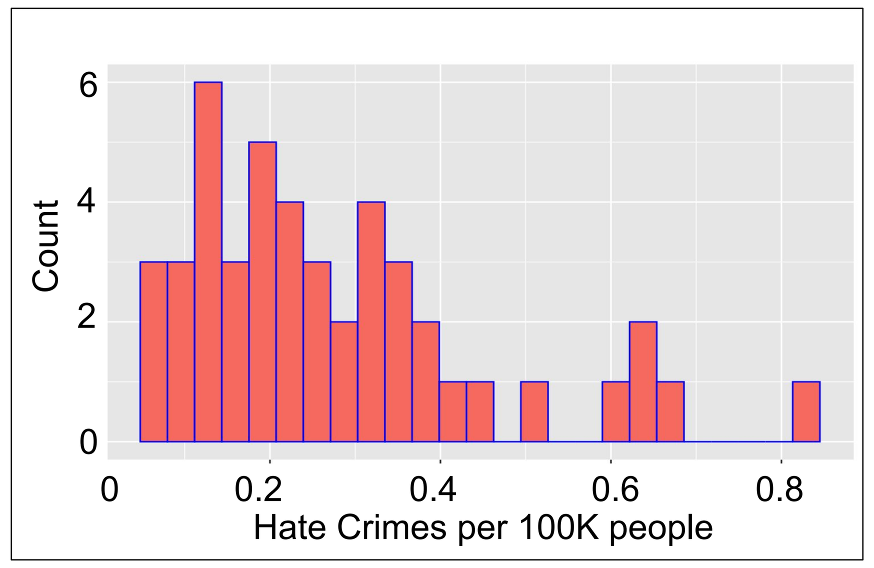


Figure 2. Distribution of response variable. Histogram of 'Hate Crimes per 100K people'. Most states observe hate crime rates of ~0.2, which is equivalent to 1 incident per 500,000 people.

Analysis

We apply the following models to our data find the association between the predictor variables and crime rates:

- Linear regression
- Best GLM (BIC, AIC)
- Regression Tree
- Random Forest
- Gradient Boosting
- K Nearest Neighbors (KNN)

Models are trained using 70% of the data. The remaining 30% of the data is used to test each model.

Analysis Summary

Best GLM (AIC) and Random Forest models generate the lowest average prediction error.

Model	Linear Regression	Best GLM (BIC)	**Best GLM (AIC)**	Regression Tree	**Random Forest**	Gradient Boosting	KNN
$\sqrt{\text{Mean-Square Error (MSE)}}$	0.195	0.336	**0.114**	0.176	**0.130**	0.192	0.173

Table 3. Square root of MSE values corresponding to each model. The square root of the MSE value corresponds to the average prediction error of the model. A lower value indicates more predictive power.

The plot below illustrates the quality of fit for Best GLM AIC and Random Forest. The models appear to overpredict hate crimes for states with lower crime rates.

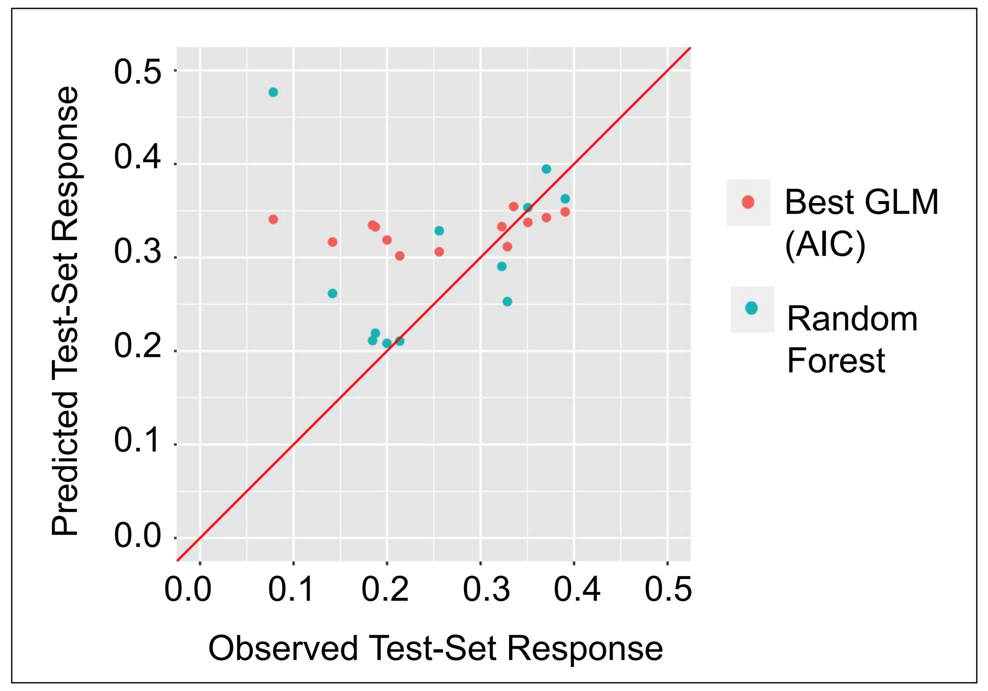


Figure 3. Overlay of Random Forest and Best GLM (AIC) quality of fit to data. Plot represents the observed test-set response values versus the predicted response values produced from each model. Points that fall along the red x=y line indicate that the model has high predictive power for this data.

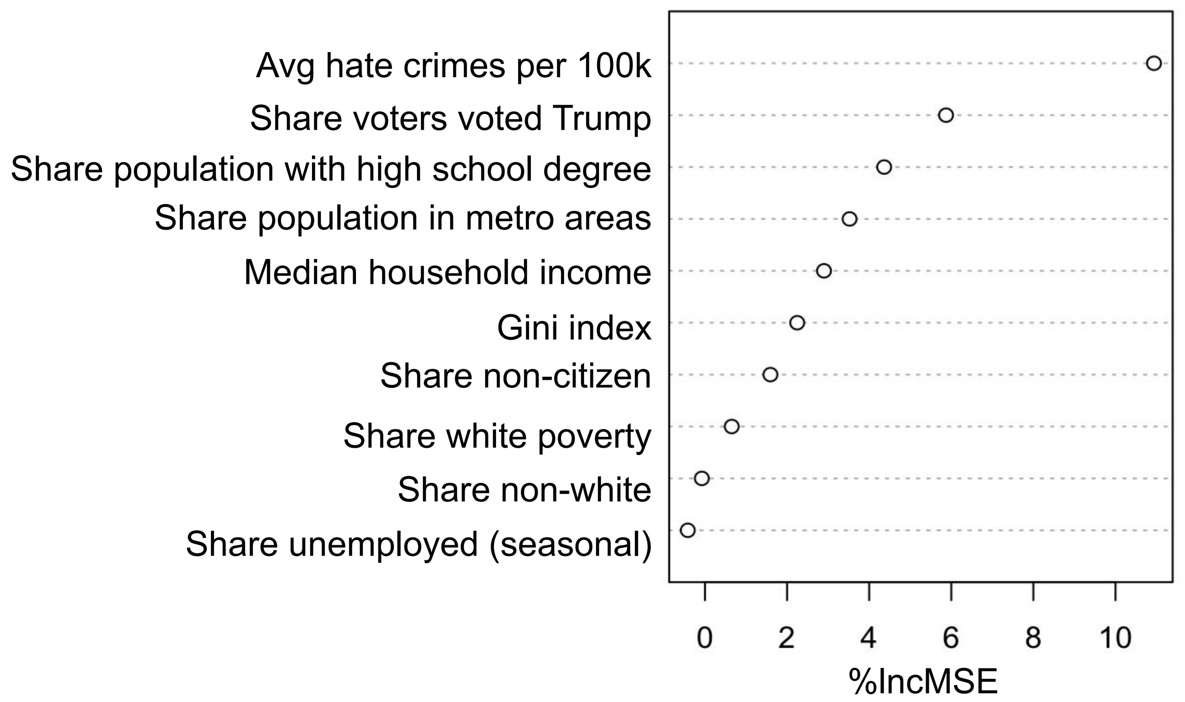


Figure 4. Random forest variable importance plot. Plot illustrates the most significant predictor variables in descending order.

The plot above indicates the most important predictor variable for this data: 'Average hate crimes per 100k' between 2010-2015, which is the reported crime rate per state prior to the 2016 election.

Predictor variables are arranged in the plot by decreasing importance. Higher %IncMSE indicates greater importance for the model.

Coefficients	**Share of Population with High School Degree**	**Gini Index**
Estimate	0.543	0.326
Std. Error	0.209	0.212
t-value	2.598	1.536
Pr(>|t|)	***0.0148**	**0.136**

Table 4. Best GLM (AIC) summary output. This table shows the model coefficients that are retained from the AIC model and its residual standard error and p-value.

The R^2 value represents that amount of variance explained by the model out of the total variance in the data. Higher R^2 indicates a better fit. Although AIC improves the square root of MSE, its R^2 is lower compared to linear regression:

- R^2, linear regression = 0.423
- R^2, best GLM (AIC) = 0.148

Conclusion

- Best GLM (AIC) and Random Forest produced the lowest average prediction error.
  - Hate crime rates reported prior to the 2016 election are most predictive for post-2016 election hate crime rates by random forest (Figure 4)
  - Interestingly, best GLM (AIC) retained different coefficients for its model compared to random forest: 'Share of population with high school degree' is the most predictive variable (Table 4)
- This analysis indicates that different models weigh varying importance for the predictor variables.
- However, due to lack of data, we were unable to generate a model tailored for all 50 states.

References

[1] Majumder, M. (2017, January 23). *Higher Rates of Hate Crimes Are Tied to Income Inequality.* GitHub.

[2] Reinhart, A. (2017, February 24). *Hate crimes after the 2016 presidential election.* Nifty Datasets.

[3] Freeman, P. (2021, December 14). *Statistical-Learning-Overview.* Github.