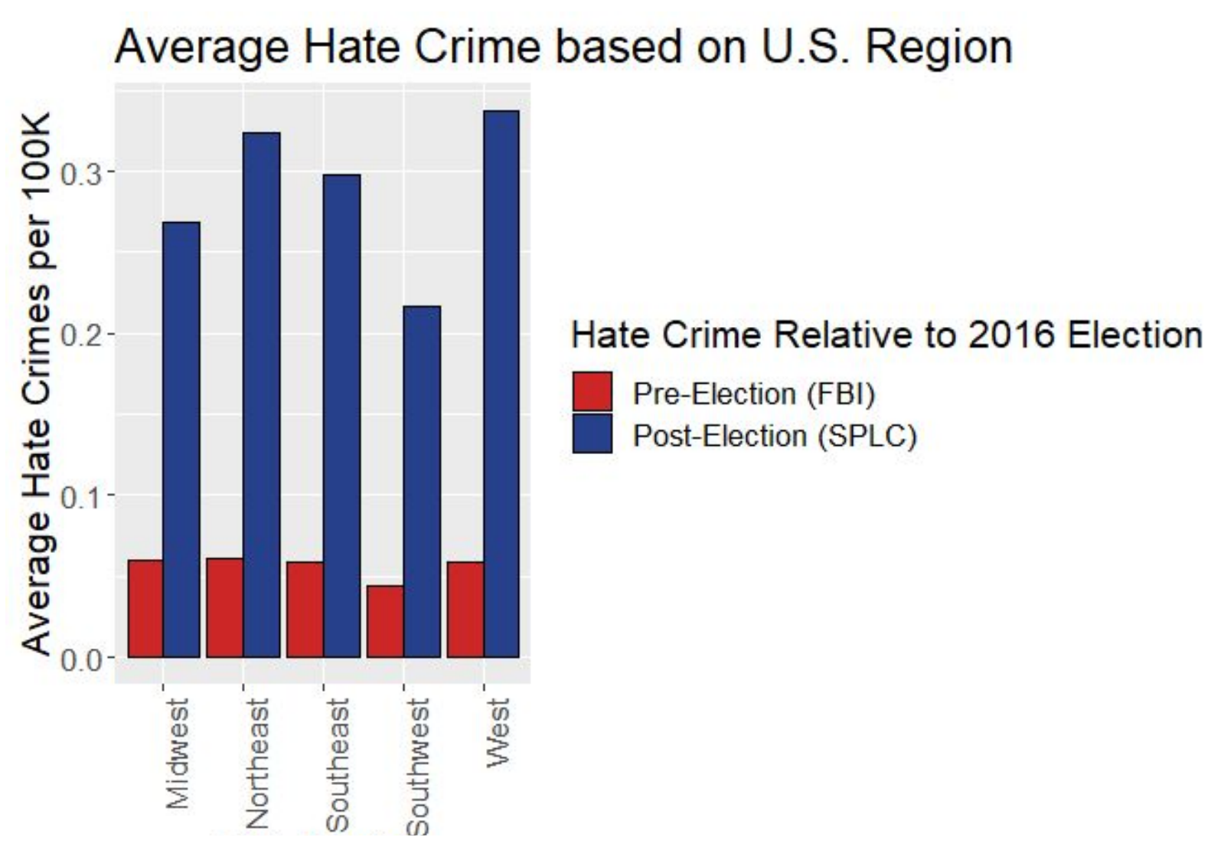
Utilizing Machine Learning Methods for Hate Crime Prediction in the USA

By: James Daley, Joon Jung, Khidong Kim, Srinidhi Manikantan, Heathvonn Styles

Background & Introduction

After the 2016 U.S. presidential election, a spike in reported hate crimes was reported by the media. The news website FiveThirtyEight compiled hate crime data from two sources. First, hate crime data were reported by the FBI from 2010-2015. Second, hate incidents (events not as severe as hate crimes) data were reported by the Southern Poverty Law Center from November 9-18, 2016 (immediately after the election).

After the election, the rate of hate crimes increased five-fold across the country. Although the SPLC recording incidents includes more events than just crimes (since not every incident is reported to the police, and the police are not required to report to a centralized database), both datasets agree that the most hate crimes are committed in the West and Northeast.



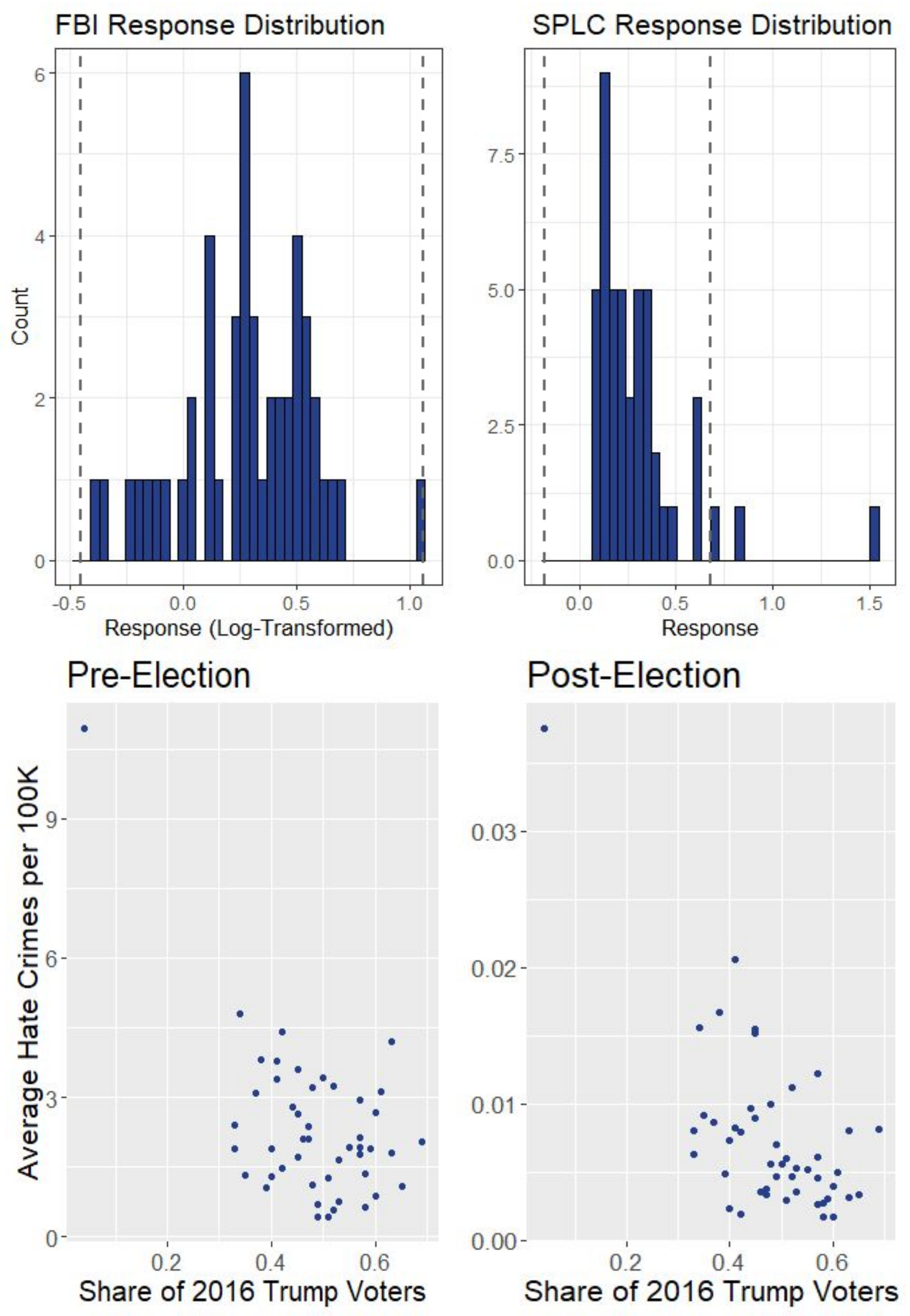
The objective of the study is to build the best predictive models for both the FBI data and SPLC data. The predictor variables include socioeconomic demographics such as income inequality, unemployment, education, poverty, and political opinion.

Data Pre-Processing

The dataset contains 51 observations (50 U.S. States and District of Columbia) and 12 variables (11 numerical and 1 categorical). Two response variables `avg_hatecrimes_per_100k_fb` (pre-election) and `hate_crimes_per_100k_splc` (post-election) were identified and analysed separately.

Given the limited data available, the intention was to preserve as much data as possible, without affecting the model significantly. Missing values identified in the response variables were removed (i.e., 4 rows) and those identified in the predictor variables were imputed with the median value (i.e., 2 entries). Any datum outside the IQR of the response variables were considered important outliers. The `state` predictor contains unique string entries, so it does not affect the regression analysis and was dropped.

In order to optimize model performance, the FBI Response variable was log-transformed, while the SPLC Response variable was not. Additionally, no significant associations were observed between any predictor and response variables.



Methods

The original dataset was randomly sampled and split such that 70% of the data was in the training set and the remaining 30% was kept in the testing set.

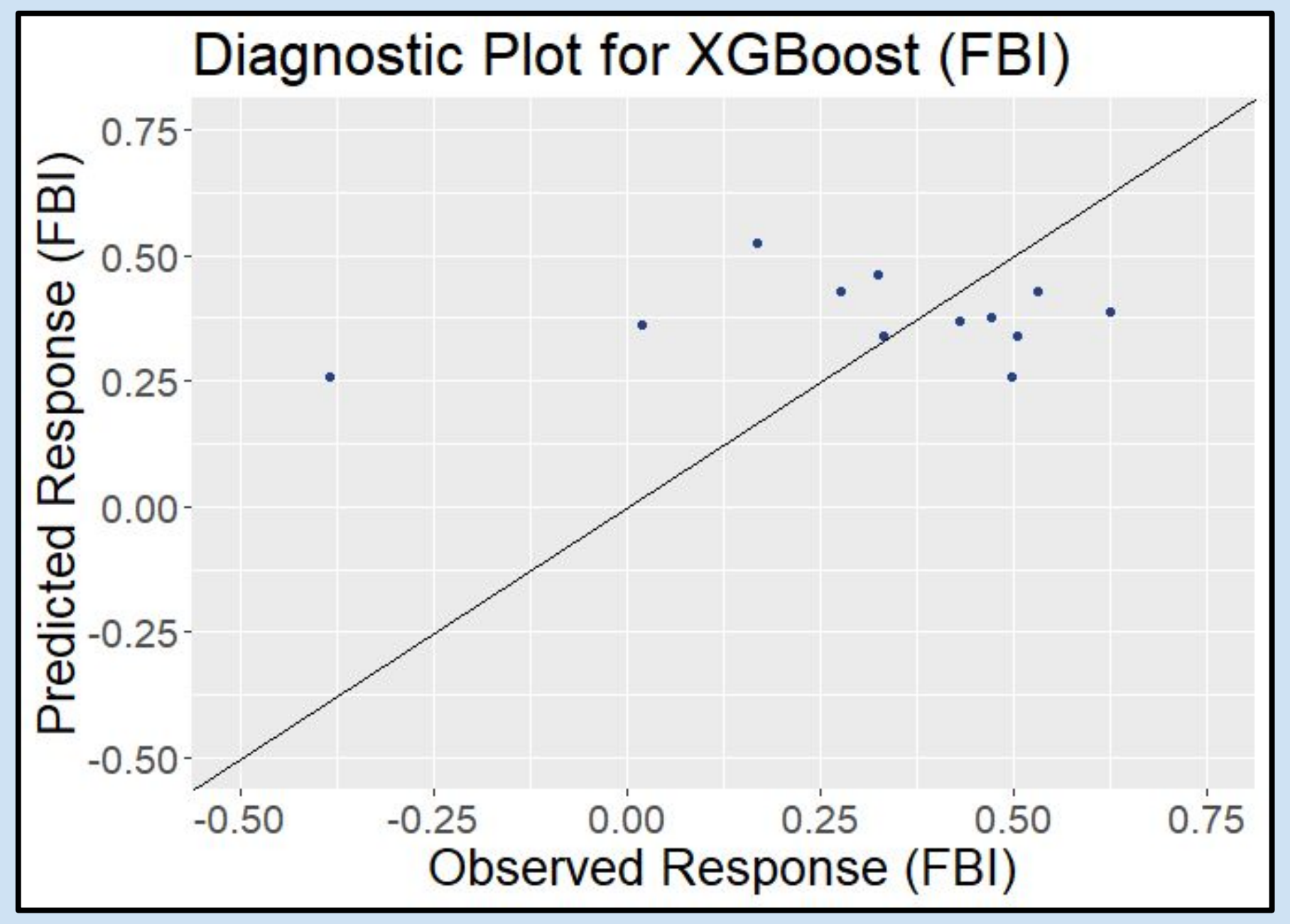
Since the two target variables identified are both continuous, different regression models was used for our prediction, namely:

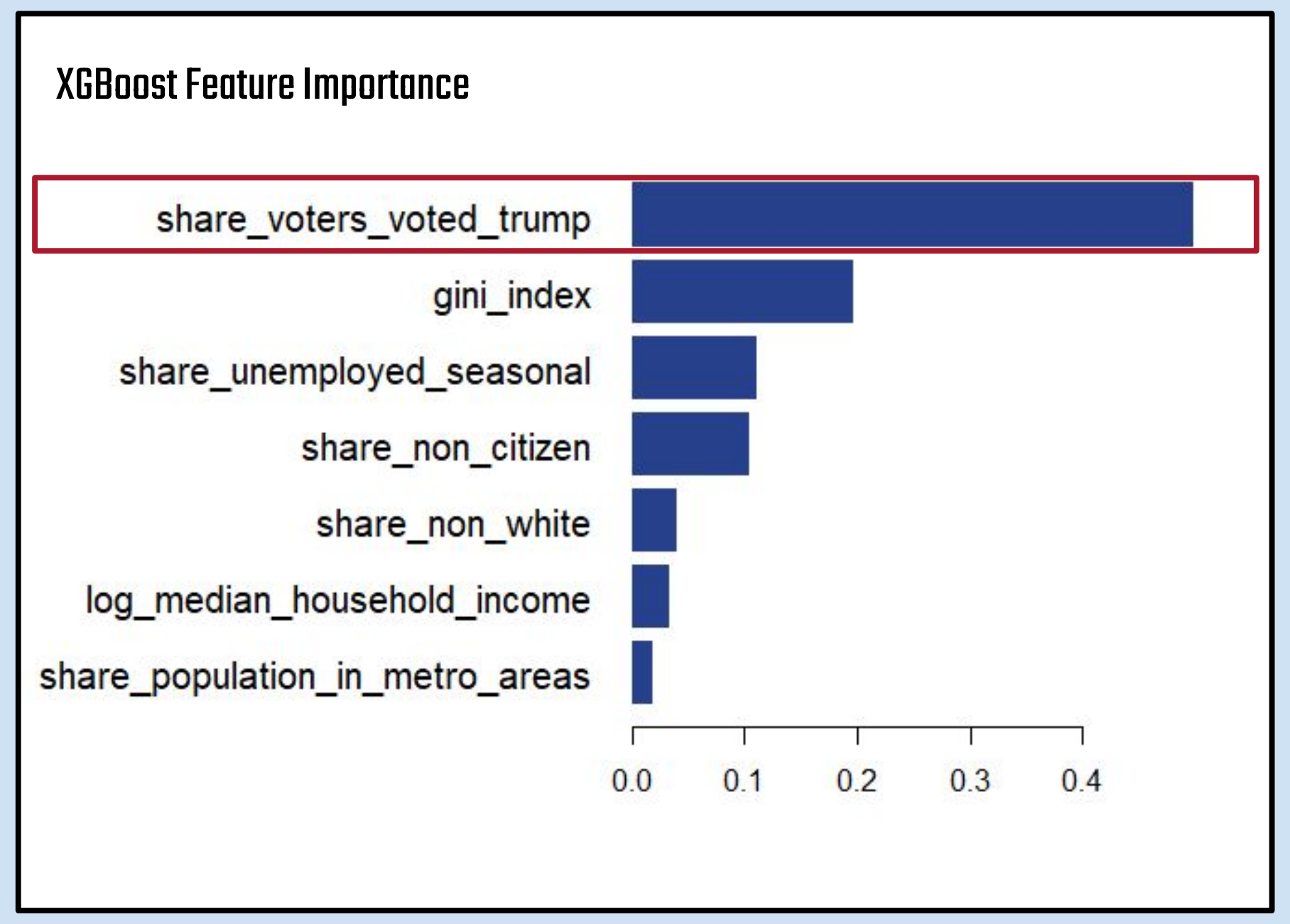
- Linear regression
- Best Subset Selection (BSS) using Bayes information criterion (BIC) and Akaike information criterion (AIC) as model selection criterias
- Regression Tree
- Random Forest
- Extra Gradient Boosting (XGBoost)
- K-Nearest Neighbors (KNN)
- Support-Vector Machine (SVM) using Linear, Polynomial and Radial Kernels

The metric used to evaluate the model performance for prediction was "Mean Squared Error" (MSE). The model with the lowest MSE was chosen to be the model with the best performance.

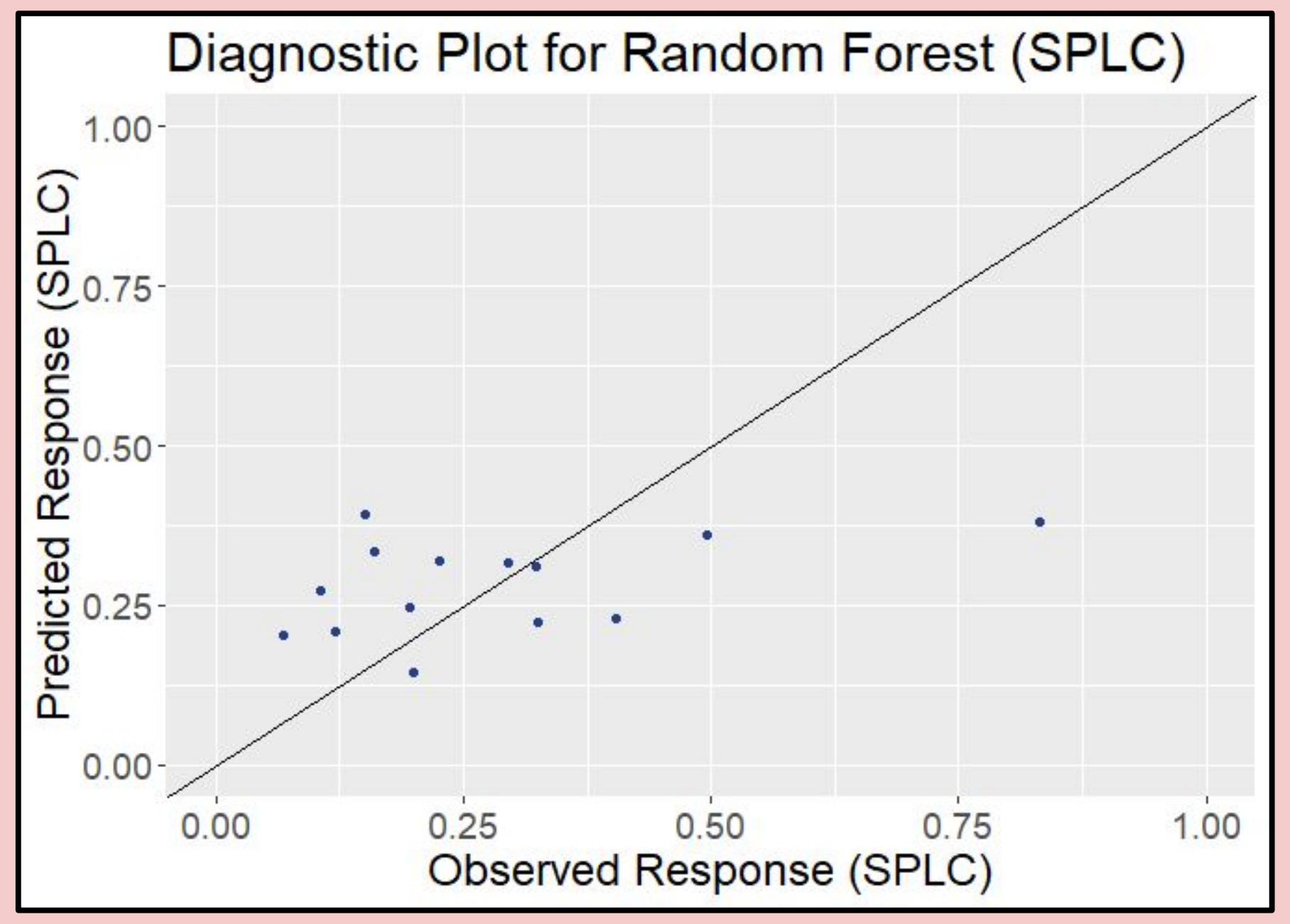
Analysis & Results

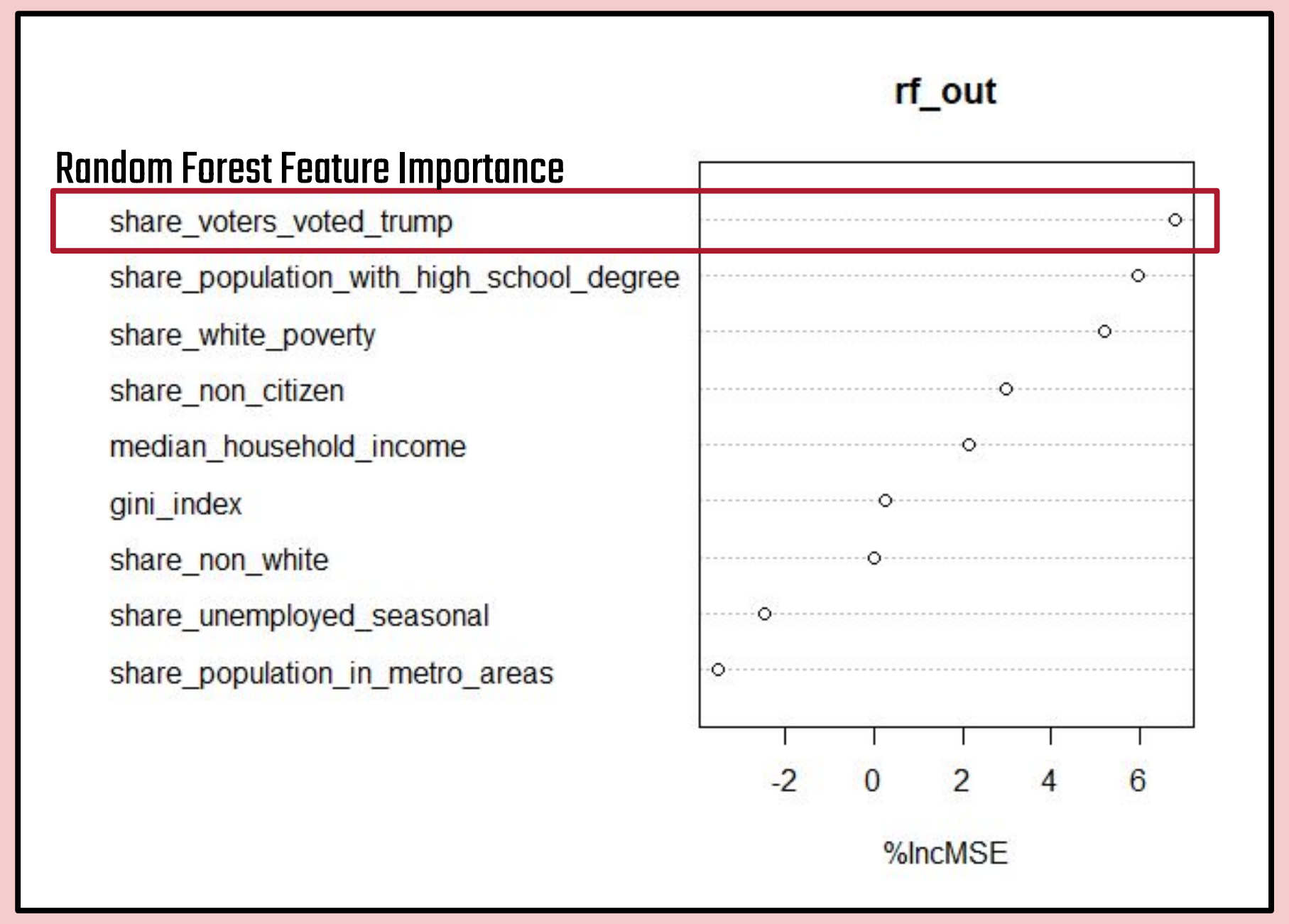
Best Model (FBI) - XGBoost





Best Model (SPLC) - Random Forest





Performance of Models on FBI and SPLC Datasets:

Models	MSE (FBI)	MSE (SPLC)
Linear Regression	0.123 (R^2 = 0.3514)	0.037 (R^2 = 0.5443)
BSS BIC	0.116	0.038
BSS AIC	0.116	0.032
Regression Tree	0.119	0.038
Random Forest	0.096	**0.030**
XGBoost	**0.072**	0.054
KNN	0.093	0.038
SVM (All Kernals)	0.102	0.038

Best Subset Selection (BSS) Feature Coefficients:

Models (Response Variable)	Important Features

BIC (FBI):	`share population in metro areas,`
	`share population with high school degree,`
	`share_gini_index,` **`share_voters_voted_trump`**

AIC (FBI):	`share population in metro_areas,`
	`share_voters_voted_trump`

BIC & AIC (SPLC):	**`share_voters_voted_trump,`**
	`share_non_white`

Conclusions

The best predicting model for the FBI dataset is XGBoost, while Random Forest is the best predicting model for the SPLC dataset. Among majority of models, `share_voters_voted_trump` was consistently identified as one of the most important variables. This is despite the fact that the variable had little to no association with the response variables. Nevertheless, based on feature importance, it showed that states that had a high share of Donald Trump voters in the 2016 U.S. Presidential Election also had a high hate crime rate.

The hate crime rates were reported most in the West and Northeast according to both the FBI and SPLC datasets. Politically, these regions tend to be more liberal. However, regions where most conservative Trump voters reside display much lower hate crime rates. Does this contrast imply that the dataset is inaccurate or unreliable? There are too many unknown factors that are at play, but further investigation could include full reporting of hate crimes from police departments and geographic breakdown into counties instead of states.

References

1. Reinhart, A. (2017, February 24). *Hate crimes after the 2016 presidential election*. Nifty Datasets. http://rosmarus.refsmmat.com/datasets/datasets/hate-crimes/

2. *United States regions*. (2023). https://education.nationalgeographic.org/resource/united-states-regions/