Police Killings

Lily Dinh (200111620), Ronak Arora (193156510),
Winnie Szeto (200553800), Por Zhang Ooi (203269350)

ST494 Project

Devan Becker

April 12, 2023

Table of Contents

Executive Summary

This report outlines our study on the “Police Killings” dataset from fivethirtyeight. The dataset was obtained from the Guardian’s database in 2015 and documents factors related to police killings. The data was analyzed, cleaned, and transformed before using to create the models. Random Forests, Decision Trees, and Neural Networks were used for modelling and Principal Component Analysis and K-Means clustering (elbow method) to clustering and data analysis.

Predictive variables like age, gender, race ethnicity, and income are in the analysis to examine the police killings data. Using these variables for our models, the Decision Tree provides the highest accuracy and Neural Network was the lowest. The Decision Tree also outperforms the other two in terms of RMSE. It is noted that Neural Network was added in as a later model to create a better comparison between the models.

Introduction

Starting in 2015, the Guardian started a database to track Americans killed by police during the year. The data was created as the official statistics have been flawed, so the Guardian’s set pulls information from media coverage, submissions, and other open sources while conducting their own verifications. This set includes demographic information of those killed and it is evident that the killings are most prominent in lower-income and Black communities, which we believe will be presented as a pattern through our models.

Motivation

For this study, our group will be looking at prediction to identify similarities within the groups and possible factors that are more important than the others that resulted in a police killing. More specifically, we are looking at the factor of the cause. To do so, we will be using Random Forest, Decision Tree Modelling, and Neural Networks techniques. Our target audience would be the general public as a source of information and could be useful to different police departments to identify any areas of concern.

Data

The data we will be using for our modelling approaches were retrieved from the fivethirtyeight repository on GitHub, which was obtained for the story "Where Police Have Killed Americans in 2015" combined with Guardian data. It documents variables based on different incidents of police killings from the census data calculated from 2015.

Before we used the data in our models, we performed exploratory data analysis to help clean our set and combine any similar factors. First, we cleaned our data by removing any empty entries or columns with unknown entries. Then, with unknown entries removed, we scaled all the data that are of numeric values and changed the data type of some categorical attributes to numeric.

```
# Convert Gender 1 -> Female, 2 -> Male
dataFD$gender <- as.numeric(factor(dataFD$gender))
unique(dataFD$gender)

# Convert raceethnicity [1]Black [2]White [3]Hispanic/Latino [4]Unknown [5]Asian/Pacific Islander [6]Native American
dataFD$raceethnicity <- as.numeric(factor(dataFD$raceethnicity))
unique(dataFD$raceethnicity)

# Convert city
# 306 Levels:  ...
dataFD$city <- as.numeric(factor(dataFD$city))
unique(dataFD$city)

# Convert state
# 47 Levels: AK[1] AL AR AZ CA CO CT DC DE FL GA ... WY[47]
dataFD$state <- as.numeric(factor(dataFD$state))
unique(dataFD$state)

# Convert cause -> [1]Death in custody [2]Gunshot [3]Struck by vehicle [4]Taser [5]Unknown
dataFD$cause <- as.numeric(factor(dataFD$cause))
unique(dataFD$cause)

# Convert armed -> [1]Disputed [2]Firearm [3]Knife [4]No [5]Non-lethal [6]firearm [7]Other [8]Unknown [9]Vehicle
dataFD$armed <- as.numeric(factor(dataFD$armed))
unique(dataFD$armed)
```

Figure 1: Data Conversion Code

```
# There are still 4 factor attribute but contains numeric data, let's fix them (pov, share_white, share_black, share_hispanic, p_income
str(dataFS)
sapply(dataFS, class)

# Care pov attribute, some problems while converting here, CONVERT pov to NUMERIC here;
dataFS <- transform(dataFS, class=as.numeric(as.character(dataFS$pov)))
# Then remove OLD POV attribute
dataFS$pov <- NULL
# Rename new Pov again;
colnames(dataFS)[20] <- "pov"

# Now we can Convert All Directly
indx <- sapply(dataFS, is.factor)
dataFS[indx] <- sapply(dataFS[indx], function(x) as.numeric(as.character(x)))
convert_chr_to_num <- function(df) {
  # Find character columns
  indx <- sapply(df, is.character)
  # Convert character columns to numeric
  df[indx] <- lapply(df[indx], function(x) as.numeric(as.character(x)))
  # Return the modified data frame
  return(df)
}
dataFS <- convert_chr_to_num(dataFS)
```

Figure 2: Data Conversion Code

After the transformations to our dataset and before creating the models, we produced a heatmap to clearly identify any high correlating factors, to help us better decide what predictions we will have. With, we further cleaned our data by dropping columns that hold zero correlation.

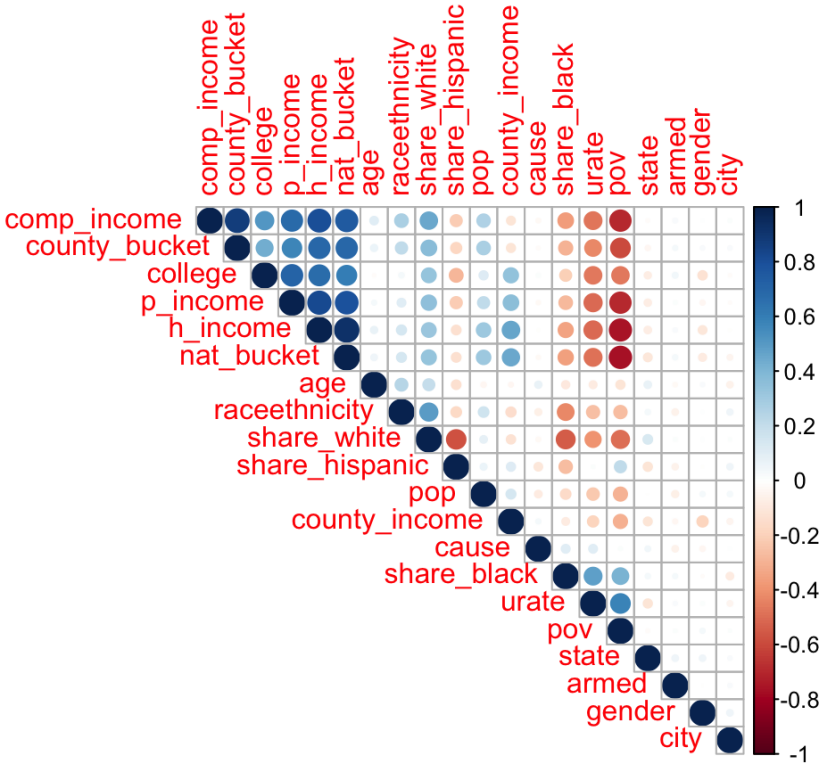

Figure 3: Cleaned Heat Map

The heatmap above is used to identify the correlation between each variable.

Models

Random Forest

This model builds multiple decision trees based on selected subsets.

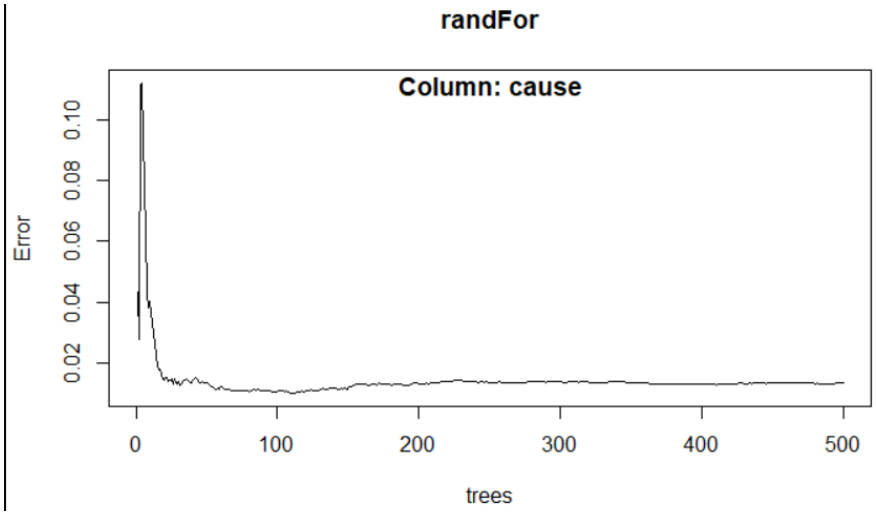

Figure 4: Random Forest Plot for Cause Factor
(Plot for all other factors can be found in the Appendix)

Recursive Forest Elimination Test

This is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. Features are ranked by the model's coef_ or feature_importances_ attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model. The graph below shows the RMSE of each variable in the model, and the fifth variable, which is pov (tract-level of poverty rate) has the highest RMSE.

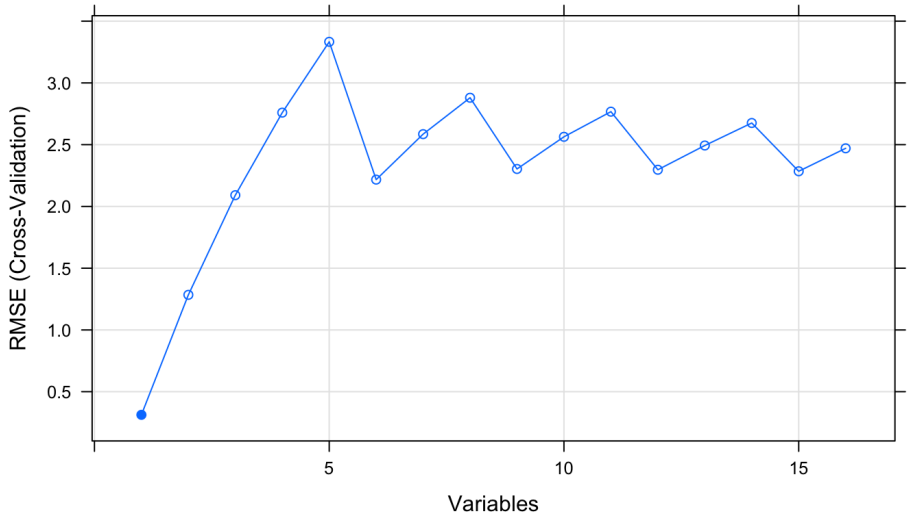

Figure 5: RFE Results Plot

Principal Component Analysis

PCA is used to reduce the dimensionality of large data sets. Allows the data to be easily visualized and analyzed.

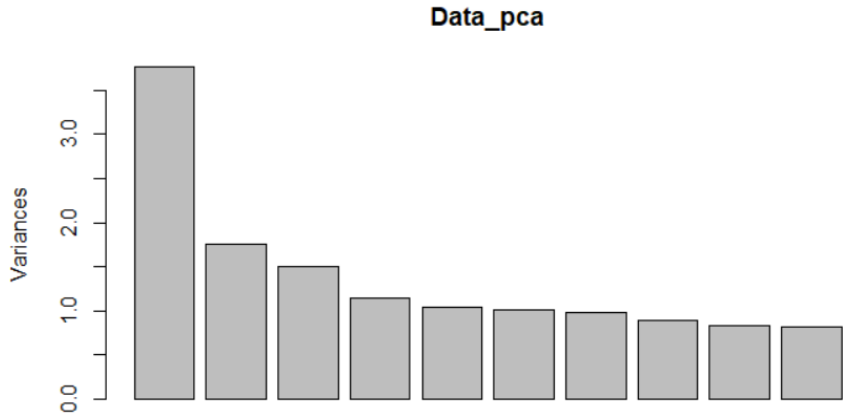

Figure 6: PCA Plot

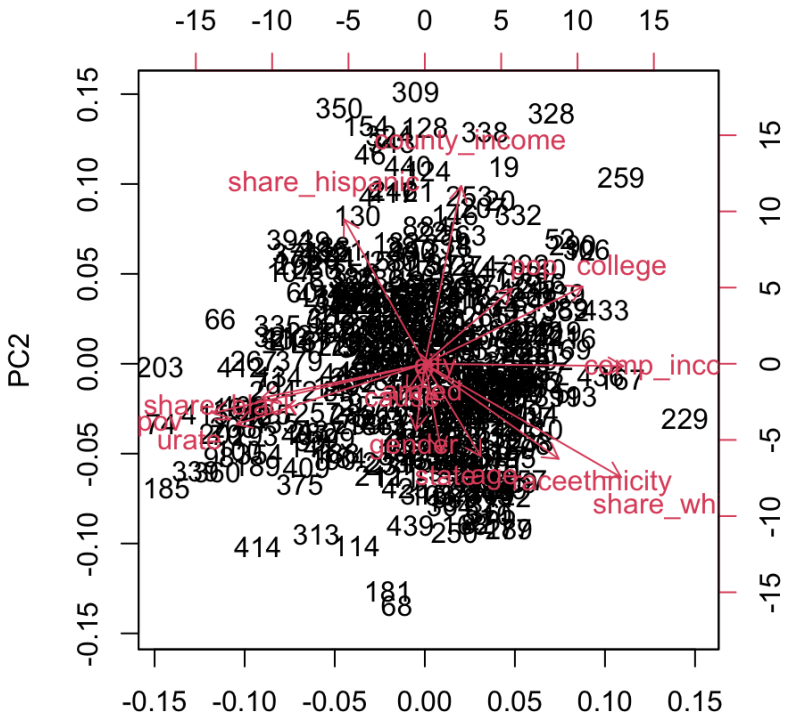

Figure 7: PCA Biplot

Elbow Method

The Elbow method is used to find the optimum number of clusters.

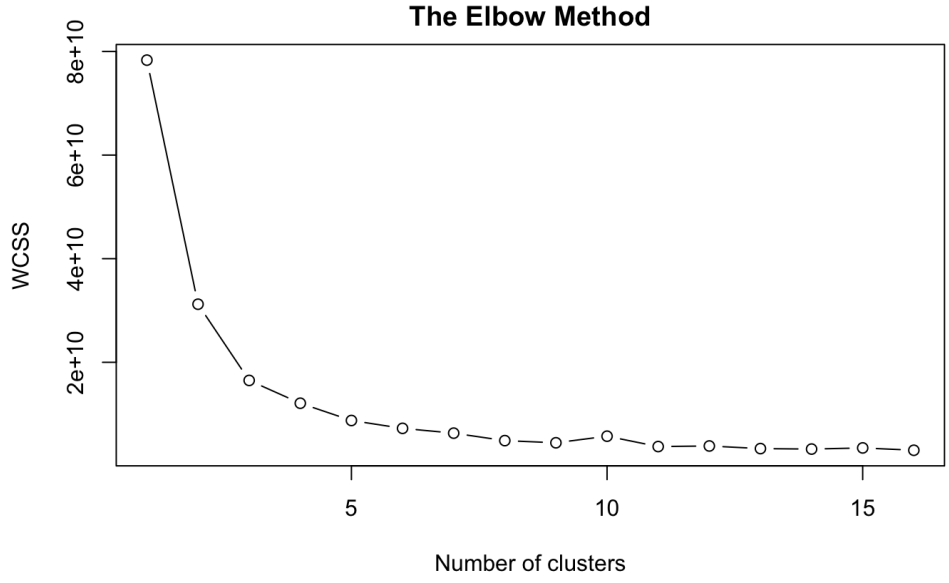

Figure 8: Elbow Method Graph

K-Means

K-Means groups similar kinds of items in the form of clusters to use for modelling.

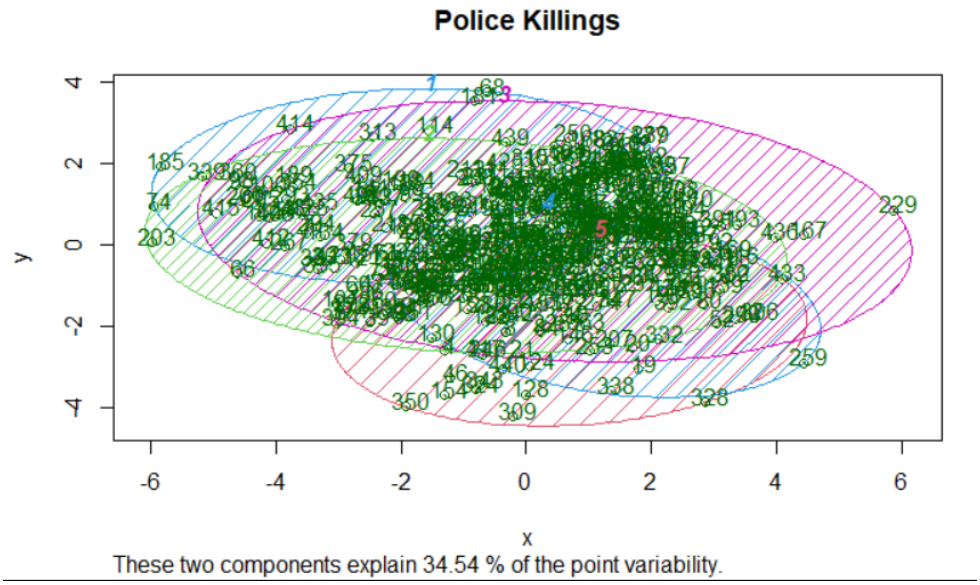

Figure 9: K-Means Clusters (5)

Decision Tree

A predictive model that selects the best feature and splits the data based on it.

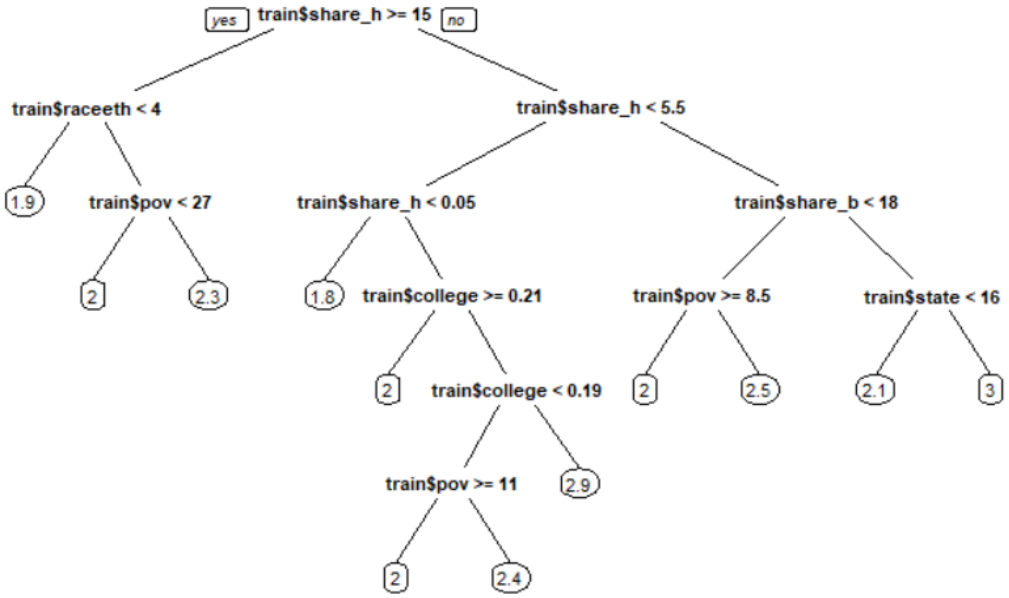

Figure 10: Decision Tree

Model Evaluation

We assigned 75% of the data to the training set and the remaining 25% to the test set. Three models were built using Decision Trees, Random Forests, and Neural Networks to evaluate the accuracy of each model. The results are shown below:

Model	RMSE	R2	Accuracy
Decision Tree	0.5245682	0.4851159	0.8545455
Random Forest	0.6245119	0.004370825	0.6909091
Neural Network	0.6758601	0.07787084	0

Based on the table, the decision tree model has the highest accuracy of 0.8545, and followed by the random forest model with an accuracy of 0.6909, and the neural network has the lowest accuracy of 0. It is not possible that the Neural Network model has 0 accuracy, more tunings are needed in the future to predict the data with this model. In terms of RMSE, the decision tree model also outperformed the other two models with RMSE of 0.5246, followed by the random forest with RMSE of 0.6245 and the neural network with RMSE of 0.6759. The R2 value for the decision tree model was 0.4851, which indicates that the model can explain 48.51% of the variance in the data. While the R2 for the other two models were lower, this means those models have poor explanatory power. Overall, the decision tree model appears to be the best performing model in terms of accuracy and RMSE.

Conclusion and Results

In this project, we developed three different models, which are Decision Trees, Random Forests, and Neural Networks, to predict the cause of death in police killings based on several features

such as age, gender, race ethnicity, and geographical location. After analyzing and evaluating the performance of these models, we found that the decision tree model performed the best with RMSE of 0.5245682, R2 of 0.4851159, and an accuracy of 0.8545455. The decision tree has the same accuracy as the random forest but has the lowest RMSe compared with the other two models. However, there is still room for improvement in the models, for example, more variables could be added to the analysis to improve the accuracy of predicting the models.

This analysis highlights the potential of machine learning techniques in predicting the cause of death of victims in police killings. The predictive variables such as age, gender, race ethnicity, and income should be included in the analysis to examine police killings data, and it would help people to have a better understanding of social issues such as police violence and to inform policy decisions that can address these issues.
Additionally, the more recent data is preferred to increase the accuracy of the data because the dataset we used has eight years difference from now.

Overall, the decision tree model is the best so far to predict the cause of death in police killings compared with random forest and neural networks, based on the available dataset. As the accuracy of the three models is not high enough to predict, further research can be conducted to improve the performance of these models or explore other models that could provide more accurate results. These findings could be useful in helping law enforcement agencies and policymakers to take necessary measures to prevent police killings.

Appendix

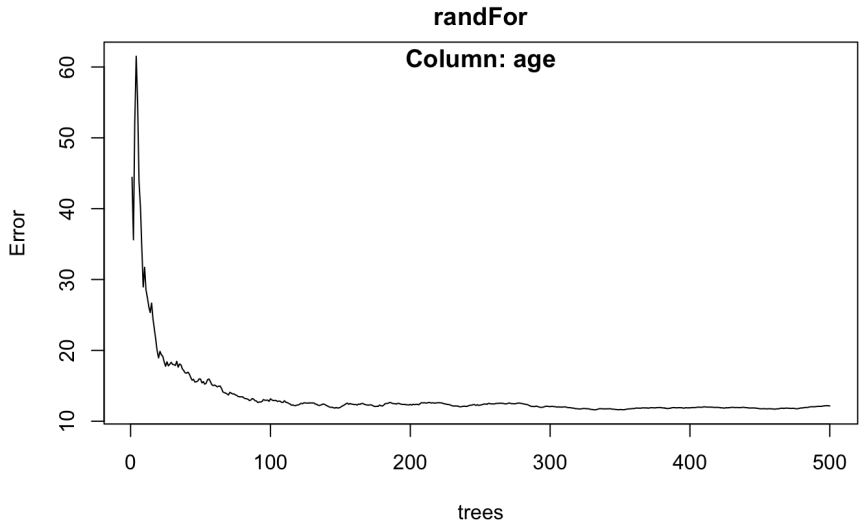

Figure A1: Random Forest Plot for Age Factor

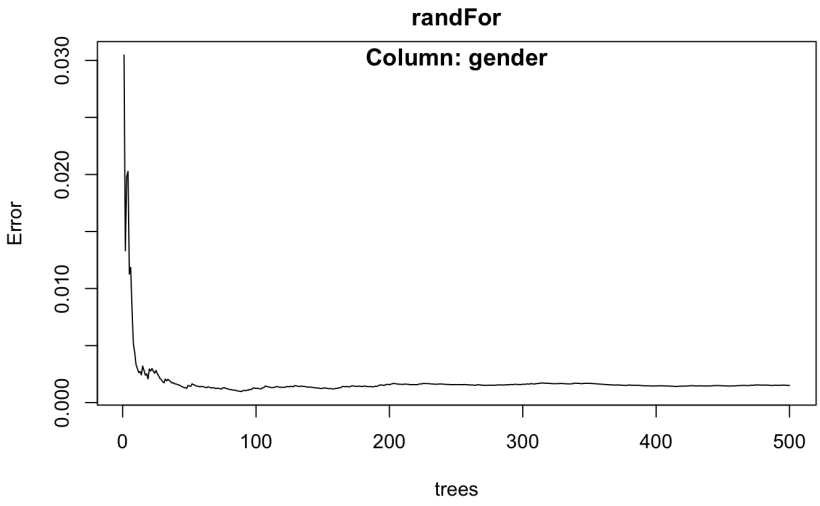

Figure A2: Random Forest Plot for Gender Factor

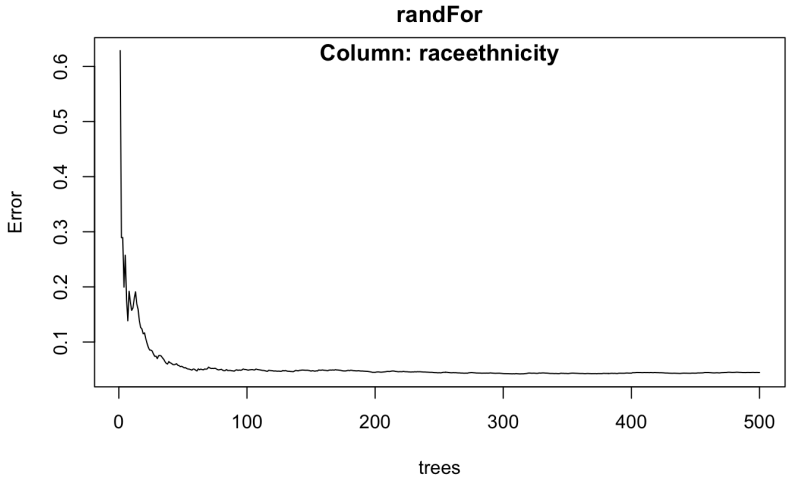

Figure A3: Random Forest Plot for Race Ethnicity Factor

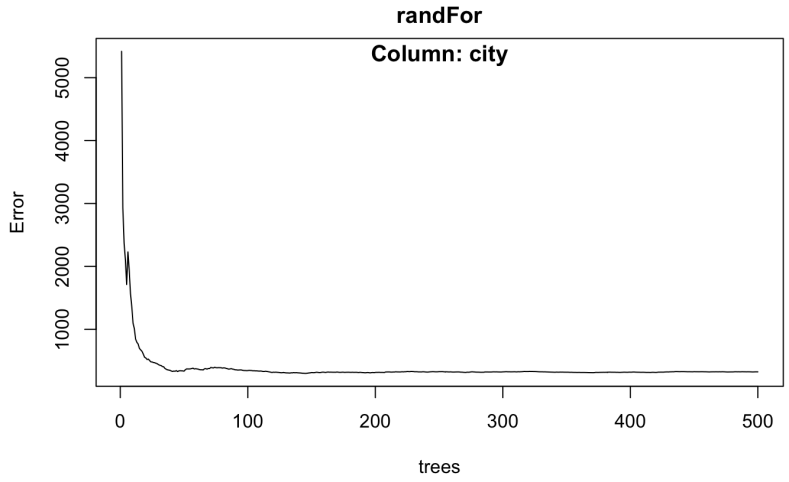

Figure A4: Random Forest Plot for City Factor

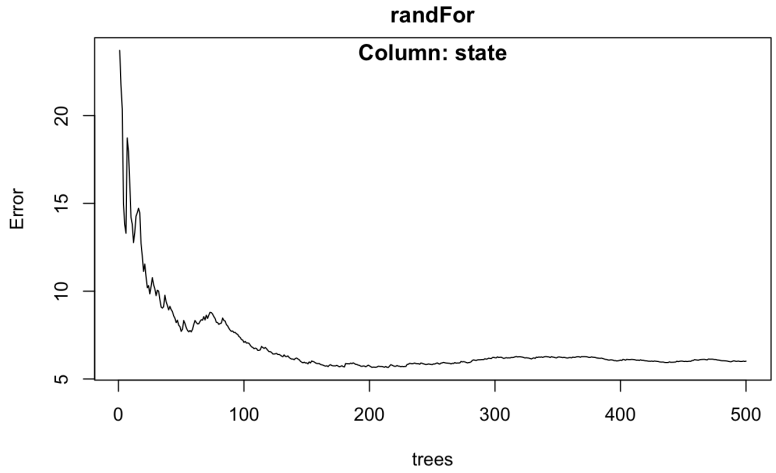

Figure A5: Random Forest Plot for State Factor

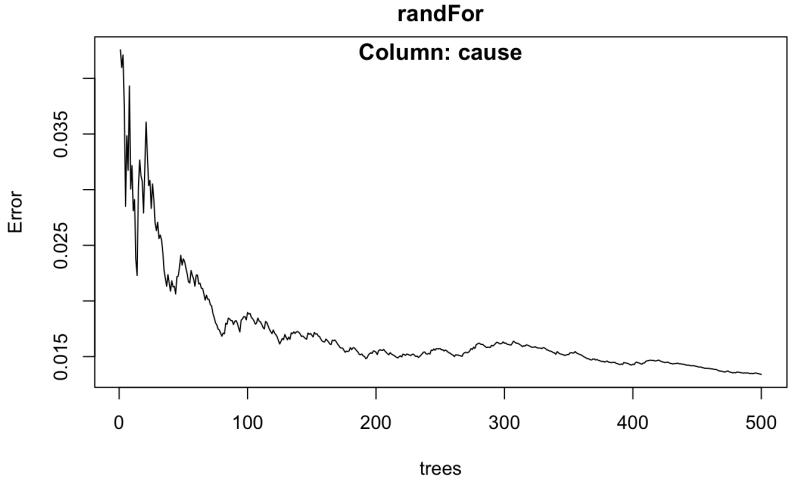

Figure A6: Random Forest Plot for Cause Factor

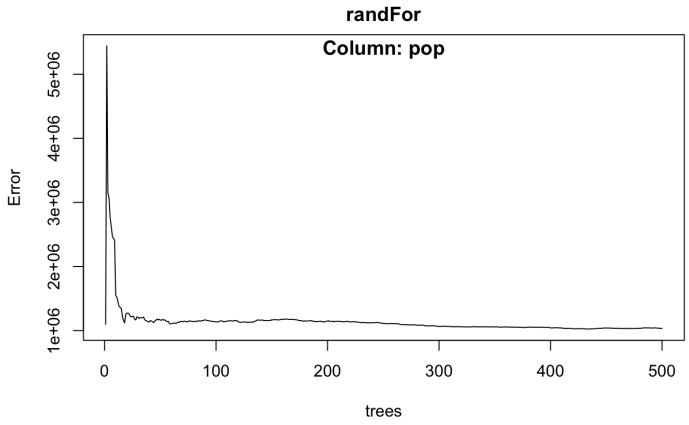

Figure A7: Random Forest Plot for Population Factor

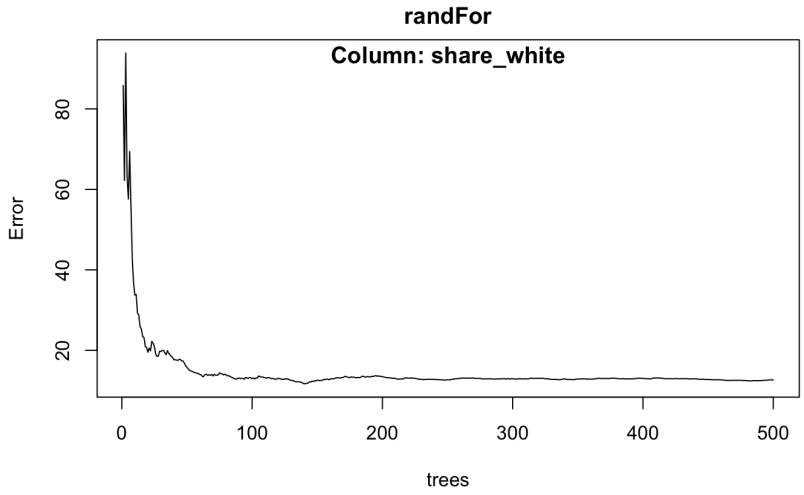

Figure A8: Random Forest Plot for non-Hispanic white Population Factor

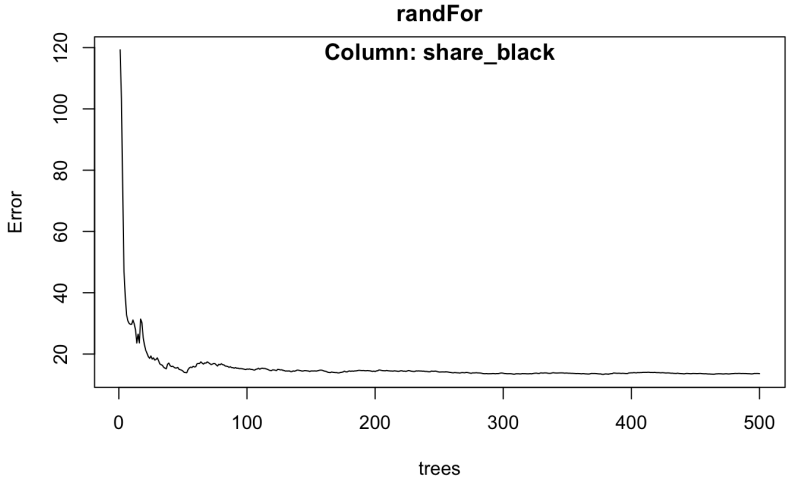

Figure A9: Random Forest Plot for black Population Factor

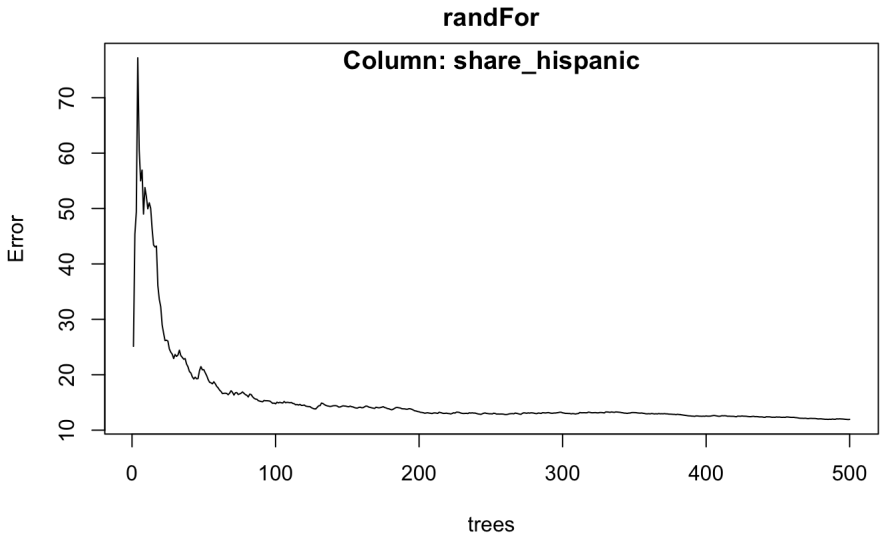

Figure A10: Random Forest Plot for Hispanic/Latino Population Factor

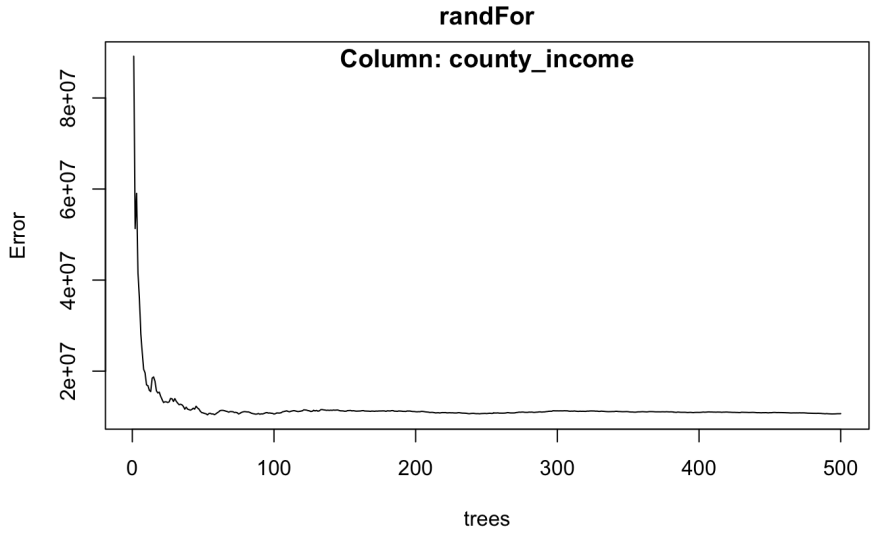

Figure A11: Random Forest Plot for County Income Factor

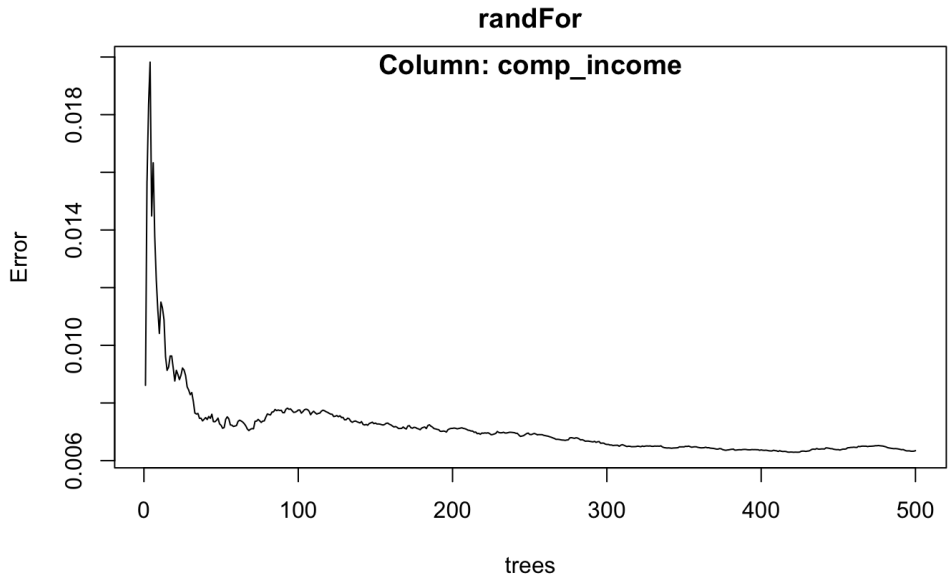

Figure A12: Random Forest Plot for County-Level Median Household Income Factor

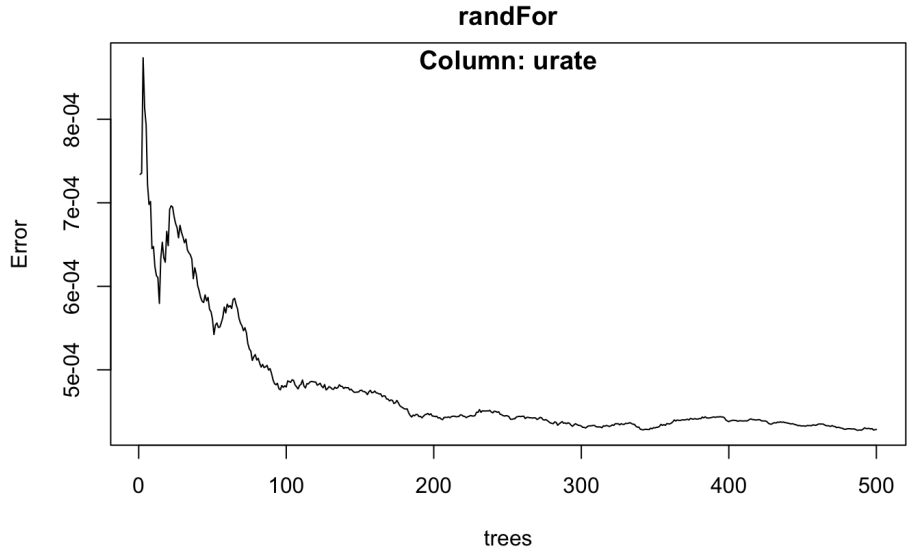

Figure A13: Random Forest Plot for Unemployment Rate Factor

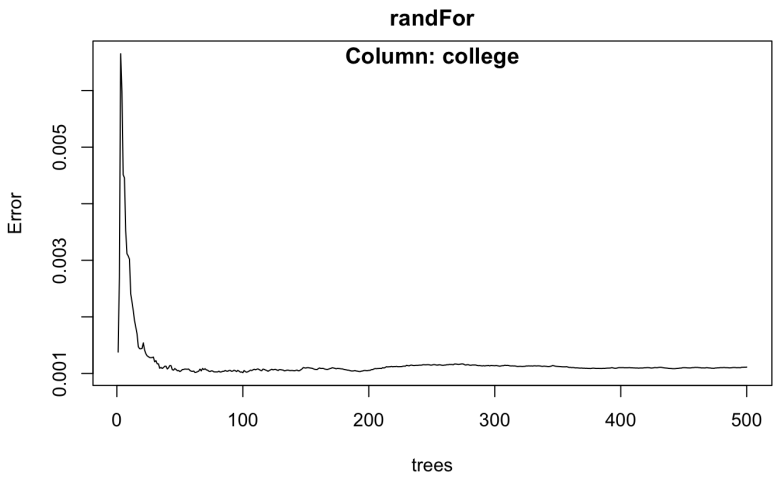

Figure A14: Random Forest Plot for College Degree Factor

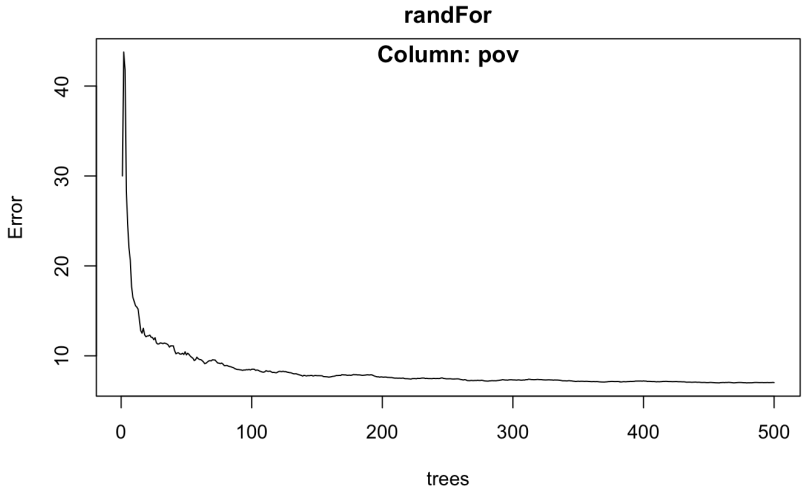

Figure A15: Random Forest Plot for Poverty Rate Factor

References

Fivethirtyeight. (n.d.). *Data/police-killings at master · fivethirtyeight/data*. GitHub. Retrieved April 12, 2023, from https://github.com/fivethirtyeight/data/tree/master/police-killings

Delineation of Work

Lily Dinh (200111620):
- Data selection and deciding the general approach to models
- Exploratory data analysis

Ronak Arora (193156510):
- Data selection and deciding the general approach to models
- Data modelling

Winnie Szeto (200553800):
- Data selection and deciding the general approach to models
- Report Outline and Writing
- Data File Cleaning

Por Zhang Ooi (203269350):
- Data selection and deciding the general approach to models
- Model Evaluation and interpretation
- Model Results