Jonah Shi

Individual Research Paper

**What are the most significant effects that influence violent crime rates across US states?**

**Introduction:**

Nowadays, there is a lot of news about shooting incidents in our lives, which brings people’s attention to violence and social safety. Furthermore, this situation is different among all the states in the US. Some of the states are safer than others; some report more violent incidents. I want to figure out what aspects have influenced the violent situation and made the situation so different among all the states in the US. The goal of this study is to explore that question using data on violent crime rates and social factors in US states from the 1970s to the 2000s.

To know the influential aspects, linear regression and random forest are used to analyze the dataset. The response variable will be violent incidents, and the predictor will be some social factors: income, population, gender, density, law, and state. For the random forest, all the data will be divided into different branches, analyzing each branch using bootstrapping and resampling to determine how much they will influence the original model. The final result will be based on how each of the experimental variables’ importance scores for violence emerges in the final model. In the end, the study will compare the RMSE(Root Mean Squared Error) and R Square between two models to find out the fitter one.

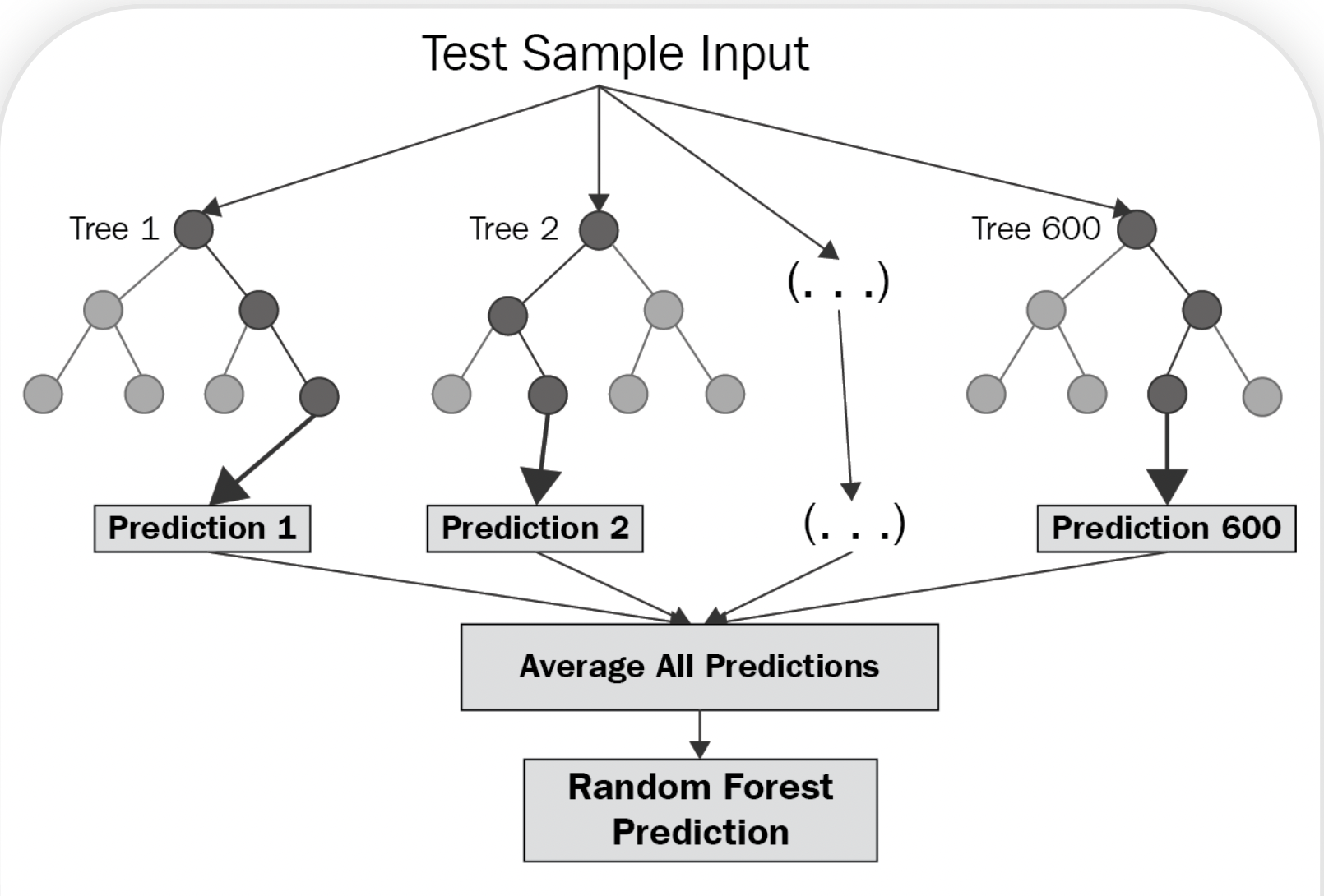
**Methods and Materials:**

The dataset used in the study is obtained from Vincent Arel Bundock’s GitHub project named gun. The data is a balanced panel of data on 50 US states, plus the District of Columbia (for a total of 51 states), by year for 1977–1999. This dataset also includes 1,174 records of each of the states’ violent crimes in different years and their state’s information: law, income, population, density, percentage of the population that is caucasian. For this study, the most important variable will be the Violent crime rate (incidents per 100,000 members of the population), which is the response variable.

The dataset also records potential candidates factors that influence violent crime: law -- “does the state has a shall carry law (requires a license to carry a concealed handgun) in effect in that year” (yes or no), state -- “full name for each of the state,” income -- “the state average income,” population -- “the state population in millions,” density -- “population per square mile of land area divided by 1,000”, caucasian -- “percent of the state population that is Caucasian, ages from 10 to 64”. These variables were selected from the dataset for their potential influences on violent crime, so the study will go over each variable to see how much they influence the response variable. To get a comprehensive view of how different the variables influence the variable Violent, random forest and linear regression are selected.

The random forest does not have many assumptions. Random forest works for most of the distribution, so the study does not need to state the normality. Thus no matter whether data is skewed or multi-model, the random forest can handle them all.

After checking the assumption, the random forest will use decision trees to help with the analysis. There are two different ways of doing the random forest-based on response variables, classification and regression. Because the response variable that I am using, violent crime, is a continuous variable, I need to use regression random forest in my analysis. The graphs below show exactly how the math works.



(Fig 1) (live tv channel, n.d.)

First, the study divides all the independent data into N branches. In figure one, N equals 600, so the data is divided into 600 trees: Tree 1, Tree 2 until Tree 600. Moreover, the study bootstrapped all the sample data and decided on a different sample of features for splitting. After that, the study will create decision trees for each node, each of the trees makes its own individual prediction. The tree shape for each node in figure one can represent the decision trees that it makes. Each of the recorded data like the dark ball in Fig one will be selected into different branches based on the decision tree model. It will provide a prediction for that tree. The last steps will be based on all the predictions. The average of these predictions will have a single result, the final values for that node. In another word, the final result, the average, will be the random forest prediction.

After getting the model’s results, the study can determine how each branch’s importance of the violent crime was by providing each of the predictors’ Mean Decrease Accuracy(IncMSE). The formula is given below, and the way it calculates is analogous to accuracy-based importance, which means it shuffles the values of the out-of-bag samples. Using MSE of jth item subtracts the out-of-bag sample divides the out-of-bag sample to give a percentage of the bootstrap data’s MSE.

%IncMSE of jth item:

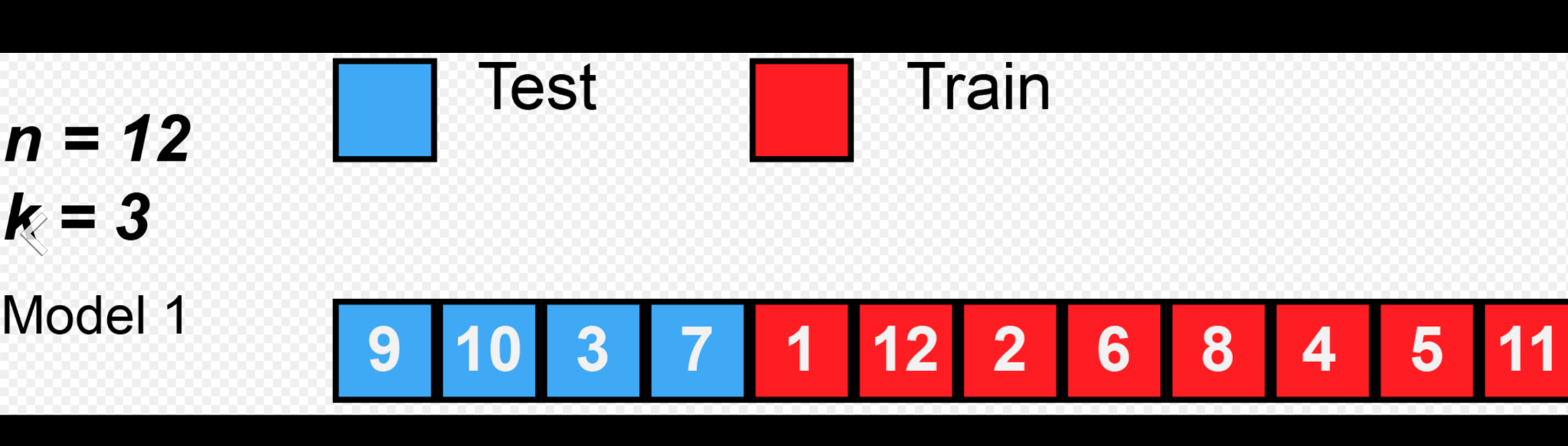
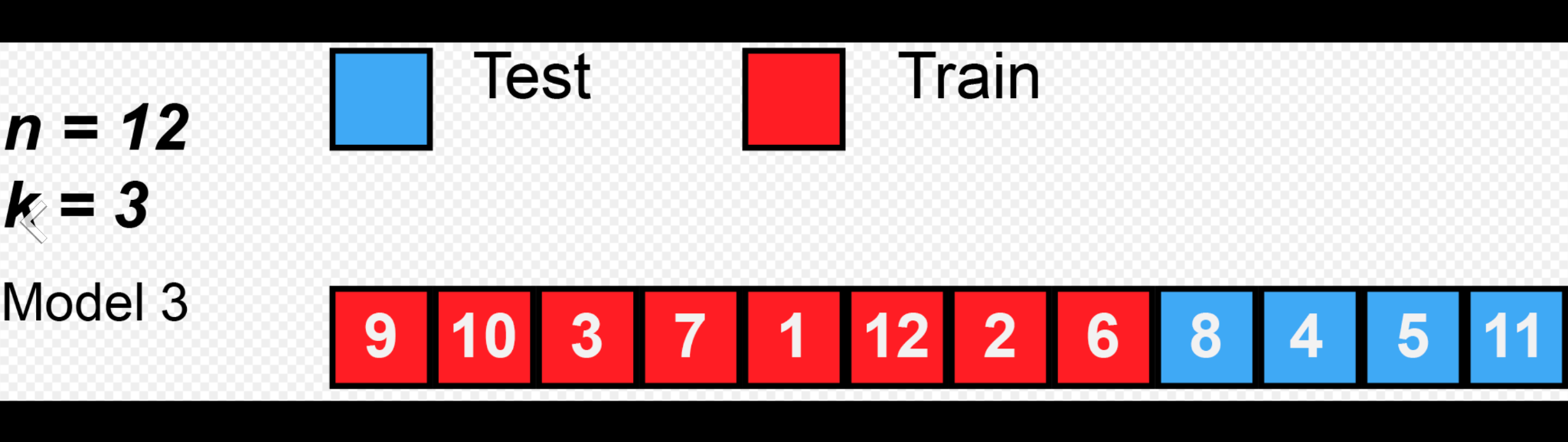
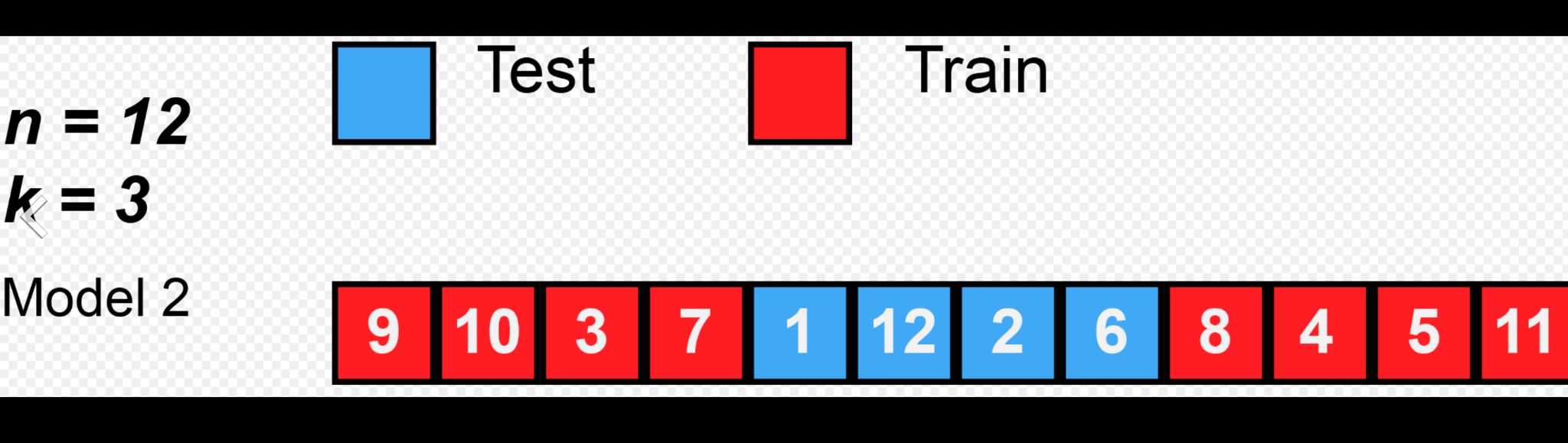
(MSE(0) is the Out of bag mean squared error, which is the mean prediction error on each training sample xi, within the data that is not be included)

IncMSE shows how much the model accuracy decreases if the study leaves that variable. If the number is higher, it will be more significant for the model, making it more likely to be included in this random forest model.

The random forest can address my research question since it provides the level of impacts for each of the predictors to the response variable. Based on the importance scores to determine whether or not the final model should include those experimental variables. In the end, it can provide me with the final models.

The last step for this study will be to compare the two models, the random forest and linear regression models, by comparing the fitness for the real data by RMSE(Root Mean Squared Error). The way of doing it is using the K-fold cross-validation. The way of calculator is by these five steps.

1. Shuffle the violence crime dataset.
2. Split the data into k groups
3. For each unique group, there will be test data, and the remaining data will be the train data. (The blue and red part in Fig two)
4. Fit model with the training set and evaluate the model with the test set.
5. Retain the evaluation scores for each of the models.

(Fig two) (*Cross-Validation (Statistics)*, n.d.)

Based on the evaluation scores, it will provide RMSE and R Squared. The research will compare the RMSE, and R Squared with both models, linear regression and random forest, to evaluate which model is better.

**Results:**

Below is a summary of each response variable record for the 1174 records.

|  |  |
| --- | --- |
| Law | State |
| Yes:285 | The number of State: 51 |
| No: 877 | The number for each State: 22 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Population (millions) | Density(population per square mile/1000) | Income(real per capita personal income in the state dollars)) | Cauc(percent of the state population that is Caucasian) |
| Average | 4.816341 | 0.3520382 | 13724.8 | 62.94543 |
| Variance | 5.252115 | 1.355472 | 2554.542 | 9.761527 |

First, I use linear regression to find the best model to represent the relationship between my response variable and predictors. I use the regression models' violent crime rate as y and all the other six variables as x to calculate the AIC(Akaike information criterion) to represent the quality of each model. The research uses the stepwise forward elimination method to get the final result model. The model will start with the empty model and add all the variables in the linear model to compare with the formal model to come out with AIC. Because AIC is also an estimator of prediction error, if AIC is smaller, the model should add the variable until all the variables are added. Otherwise, the model should not add the variables.

All the AIC data I have for all the possible models are listed below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mod | income | density +income+population+cauc | law+density +income+population+cauc | All variable |
| AIC | 16755 | 16065 | 15797 | 14176 |

Because 14176<15797<16065<16755, this means that the final model should include all the variables above, and this model will be the best fit for the conclusion.

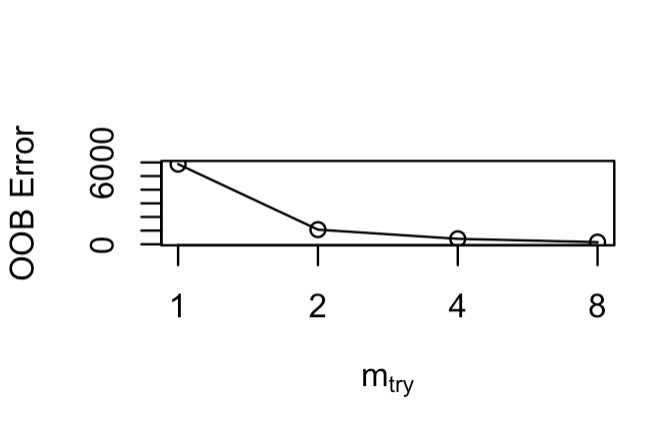
With the results shows the lowest AIC is the model

**Violent ~ Law + Population + Density + Income + Cauc + State.**

For the random forest, this research used the model with a violent crime rate as my response variable, adding all the predictors into the model to start with random forest.

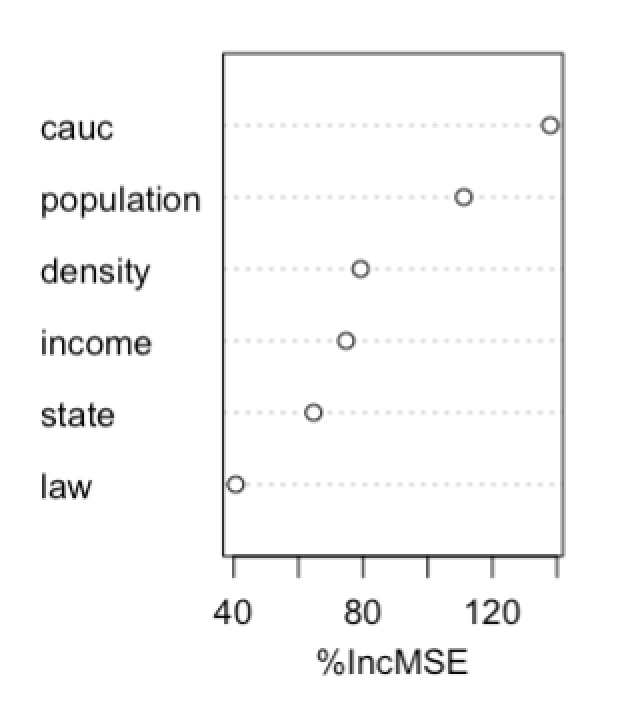
**Violent with Law, Population, Density, Income, Cauc, and State.**

For the random forest model, two things need to be decided: the number of trees to grow and the number of variables available for splitting at each tree node. For the number of trees, in R is called ntree. Because this dataset contains 1174 records, which is an extensive dataset, so the appropriate number of trees will be 500. For the number of variables for splitting, because it is a regression model, the study will use the tuneRF function in r to find out in every mtry, which will give the smallest number of the OOB error. In this study, the number 8 of splitting is the best mtry. The whole random forest means that the dataset will be divided into 500 trees and split eight times for every sub-model.



After showing the model stats, the data would be calculated the importance of each variable by using the build-in R importance function of the random forest tree’s importance level. The function is trying to calculate the whole portion of MSE minus the MSE of the out-of-bag portion of the data. Then, the research uses the same calculation with random permutations for each variable to do all the same calculations.

The average results for all trees’ differences in prediction errors are below.



The graphs show that the IncMSE(Mean Decrease Accuracy). When the percentage is higher, the variables are more likely to influence the final model. As a result, from the graph for the relationship for the influence of violent crime rate.

Caucation > population > density > income > state > law.

The data shows that law is only 40% IncMSE, which is pretty low, and the state, income, and density are at the same level, around 80. Then the population and Caucasians are the most influential predictors, which is pretty big. If the study needs to determine the predictors, I will choose the percentage of Caucasians and the population.

**Discussion:**

Based on the statistical analysis, the two methods give different regression models for each of the variables for the violence. The linear regression shows that the violence is related to the percentage of being causation, population, density, state, and law. The Random regression says that violence is related to the causation, and population.

The study compares these two methods by RMSE(Root Mean Squared Error) and R Squared.

|  |  |  |
| --- | --- | --- |
|  | Linear Regression Model | Random forest |
| RMSE | 101.4678 | 162.1804 |
| R Squared | 0.9084395 | 0.7624581 |

Larger R Squared means better fit the model, smaller RMSE means less variability and more fit for the data. For RMSE, 162.1804 is greater than 101.4678, and R Squared 0.9084395 is greater than 0.7624581. All of the evidence shows that the Linear Regression Model is better than the random forest model. Therefore, the linear regression model will be a better fit for this model.

Meanwhile, both of these two methods have three limitations. First, when researchers are analyzing, it assumes that all the variance for the violent crime is made by listed predictors. However, violent crime can also be influenced by other social events. For example, last year, the percentage of violent crime rose because of George Floyd’s event. It can increase the possibility of a violent crime ratio. In that case, the violent crime will be increased not only influenced by the predictors like income, density, population, and the percentage of being Caucasian which listed in the research. It involved other factors such as societal situations. My suggestion will be either adding more variables into the predictors or limiting in only one year and doing analysis only in that year.

Second, the dataset also contains some NA values for the law and the state, especially for Wyoming. Because of that, the results will be influenced by the missing values. In the future, if possible, the research can go deeper with these NA data, and finally, get the decision trees for these two variables for the state of Wyoming and have a better understanding of all the states in the US.

Finally, the data is only from the 1970s to the 2000s, so the study’s results may not be applicable nowadays. It will be the out-of-date data for now. The study may need more recent data to know more about violent crime recently.

**Conclusion:**  
The significant factors that will affect the probability of violent incidents happening in the U.S. from the 1970s to the 2000s are the state population in millions, population per square mile of land area, percent of the state population that is Caucasian, income(dollar), law, and state name. However, the data analysis can still be improved by collecting a more random, complete dataset.

Appendix 1:

Bootstrap:

The idea is to use observed data to see the population parameters or the distribution. Replacement for the sample from the original sample using the SAME sample size. This will create the bootstrap sample.

Appendix 2:

Linear regression result:

Call:

glm(formula = violent ~ ., data = guns)

Null deviance: 130960736 on 1172 degrees of freedom

Residual deviance: 11026698 on 1116 degrees of freedom

AIC: 14176

Call:

glm(formula = violent ~ income, data = guns)

Null deviance: 130960736 on 1172 degrees of freedom

Residual deviance: 109161939 on 1171 degrees of freedom

AIC: 16755

Call:

glm(formula = violent ~ law + density + income + population + cauc, data = gun)

Null deviance: 130960736 on 1172 degrees of freedom

Residual deviance: 47907918 on 1167 degrees of freedom

AIC: 15797

Call:

glm(formula = violent ~ density + income + cauc + law, data = gun)

Null deviance: 130960736 on 1172 degrees of freedom

Residual deviance: 59166190 on 1168 degrees of freedom

AIC: 16043

Call:

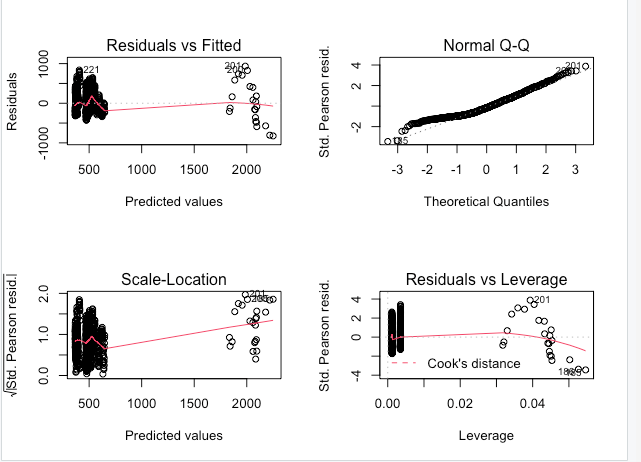
glm(formula = violent ~ law + income + cauc, data = gun)

Null deviance: 130960736 on 1172 degrees of freedom

Residual deviance: 74539125 on 1169 degrees of freedom

AIC: 16312

Normality checking



References

Arel-Bundock, v. (n.d.). *Data Sets*. Vincent Arel-Bundock's Github projects. Retrieved December 4, 2021, from <https://vincentarelbundock.github.io/Rdatasets/datasets.html>

Bhatia, N. (n.d.). *What is Out of Bag (OOB) score in Random Forest?* Towards Data Science. Retrieved December 8, 2021, from https://towardsdatascience.com/what-is-out-of-bag-oob-score-in-random-forest-a7fa23d710

Schott, M. (2019, April 25). *Random Forest Algorithm for Machine Learning | by Madison Schott | Capital One Tech*. Medium. Retrieved December 6, 2021, from https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb

Arel-Bundock, v. (n.d.). *Data Sets*. Vincent Arel-Bundock's Github projects. Retrieved December 4, 2021, from https://vincentarelbundock.github.io/Rdatasets/datasets.html

Diaz-Uriarte, R. (2007). GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC bioinformatics*, *8*(1), 1-7.

Santur, Y., Karaköse, M., & Akin, E. (2016, December). Random forest based diagnosis approach for rail fault inspection in railways. In 2016 National Conference on Electrical, Electronics and Biomedical Engineering (ELECO) (pp. 745-750). IEEE.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Lott, J. R. (2013). More guns, less crime: Understanding crime and gun control laws. University of Chicago Press.

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., De Leo, G. A., & Torricelli, P. (2011). Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. *Ecological Modelling*, *222*(8), 1471-1478

Schott, M. (2019, April 25). *Random Forest Algorithm for Machine Learning | by Madison Schott | Capital One Tech*. Medium. Retrieved December 6, 2021, from <https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb>

Cross-validation (statistics). (n.d.). Wikipedia. Retrieved December 16, 2021, from https://en.wikipedia.org/wiki/Cross-validation\_(statistics)

R code:

install.packages("randomForest")

install.packages("ISLR2")

library (tree)

library (ISLR2)

attach (Carseats)

guns=read.csv("Guns.csv",header= T)

head(guns)

mod = glm(violent~.,data = guns)

summary(mod)

mod2 = glm (violent~law+income,data = guns)

summary(mod2)

mod3=glm(violent~law+density+income+population+cauc,data= gun)

summary(mod3)

mod4=glm(violent~density+income+cauc+law,data= gun)

summary(mod4)

mod5=glm(violent~law+income+cauc,data= gun)

summary(mod5)

mod6=glm(violent~income,data= guns)

summary(mod6)

model\_summ <-summary(mod)

mean(model\_summ$deviance.resid^2)

par(mfrow=c(2,2))

plot(mod3)

library(randomForest)

guns= read.csv("Guns.csv",header=T)

summary(guns)

set.seed(100)

rf <-randomForest(violent~cauc+income+state+density+law+population,data=guns, ntree=500)

print(rf)

floor(sqrt(ncol(guns) - 1))

mtry <- m(guns[-1],guns$violent, ntreeTry=500,

stepFactor=2,improve=0.05, trace=TRUE, plot=TRUE)

best.m <- mtry[mtry[, 2] == min(mtry[, 2]), 1]

print(mtry)

print(best.m)

rf <-randomForest(violent~cauc+income+state+density+law+population,data=guns, mtry=8, importance=TRUE,ntree=500)

print(rf)

summary(rf)

#install.packages("randomForest")

library(randomForest)

#Evaluate variable importance

importance(rf)

plot(importance(rf))

varImpPlot(rf)

modelalt\_summ <-summary(rf)

mean(modalt\_summ^2)

summary(gun$law)

text(m3, pretty = 0)

library(tree)

tree.carseats <- tree (gun$violent~gun$income, data=gun)

plot (tree.carseats)

mod=(gun$violent~.)

summary (tree.carseats)

plot (tree.carseats,type = c( "uniform"))

text (tree.carseats , pretty = 0)

tree.carseats

library("lattice")

pairs(~ income + population + cauc + density, data = guns)

with(rf.cv, plot(n.var, error.cv))

install.packages("randomforest")

library("randomforest")

rf.cv <- rf.crossValidation(rf, guns[,3:7], p=0.10, n=99, ntree=500)

library("tidyverse")

library("caret")

set.seed(123)

train.control <- trainControl(method = "cv", number = 10)

# Train the model

model <- train(violent ~., data = guns, method = "lm",

trControl = train.control)

# Summarize the results

print(model)

set.seed(123)

train.control <- trainControl(method = "cv", number = 10)

# Train the model

library(rpart)

model <- train(violent ~ .,

data = guns,

"rpart",

tuneLength = 9)

# Summarize the results

print(model)