Predicting crime rates of the counties in New York State using statistical models

Aiyngaran Chockalingam, Chandan Singh, Gurupraneeth Rao, Sandeep Purushotham

Group 8

1. INTRODUCTION

In today's world of crime, violent crime rates are decreasing according to statistical data released by URC (Uniform Crime Report). According to the researcher in 2010, violent crime dropped almost 6 % (Clear, Cole & Reisig, 2013). In 2011, violent crimes rates were the lowest it had been in the past 40 years. However, we witness crime occurring everywhere and almost every day on news. The statistics on crime rates provides information on crime trends, but fails to demonstrate geographical distribution on crimes occurring around the country and what factors are influencing the most. With the help of statistical learning methods, we can predict crime rates based on past data and what factors contribute to it. Many research paper and statistical models [1-19] have been implemented to capture the relationship between crime rates and socio-economic factors. Our analysis is implemented using county-level data on crime rates with the help of population and fire arms rates, prison admissions, socio-demographics, and adult arrests during the period 2008-2017 for the state of New York. Our research is focused on identifying the key factors that affect the crime rates, comprehending how these factors influence the crime rates and implement non-linear statistical models to predict the response. Our research can be used as a framework to predict the crime occurrences and factors influencing it for other states in the US.

2. LITERATURE REVIEW

Existing research studies [1-9] on predicting crime rates takes into account the conventional factors – population, age group, income, among others which are limited and generic. The link between the dependent factors and predictors in most of the research studies conducted so far has been established by taking into account limited statistical methods which are flexible and induces variance in the model. Van B. Shaw [1] employed a correlation method – [27] statistical tool used to measure the relationship between two or more variables – between each crime rates calculated from different statistical sources and the population characteristics of all counties in the state of Minnesota. The author obtained the statistical reports of crime with respect to Judicial crime rate and offenses known to police, and performed correlation between population characteristics such as Total Population, percent of Urban Population, percent of population of age 14 and above seeking employment, and others. However, the number of parameters included in the study by the author is limited and the method implemented is only restricted to correlation factors; whereas, advanced statistical learning methods such as Linear Method, GAM, and Random Forest do better in predicting the crime rates, which has been further demonstrated from our research. While implementing any statistical model, selecting a feature which gives the best response prediction is of utmost importance. In [2], researchers used the chi-squared test to select the best features in order to implement the statistical model. Moreover, the researchers explored three approaches to predictive modelling: Linear regression, logistic regression, and gradient boosting method. The results with the predictive highest accuracy have been achieved using the gradient boosting method. There are other known feature selection methods, of which few are Best Subset Selection, Ridge & Lasso, Wrapper methods, etc. Using multiple feature selection methods to predict the response might have produced better results, which author failed to demonstrate. The researcher's

[3] study on crime rates dealt with describing what the problem with using Linear models is, which was residuals not following a normal distribution. The authors then went on to explain the probable solution, which could be transformations of data, but multi-collinearities may exist among predictors leading to controversial results. In addition to conventional methods – Linear Models, Logistic Regression, Gradient Boosting, among others – researchers have lately been using advanced methods like Deep Learning to predict the crime occurrence [6-8]. In [6] the researchers have employed Deep Neural Network (DNN) with feature-level data fusion, which resulted into 84.25% accuracy, precision of 74.35, recall of 80.55, and AUC of 0.8333 and claims to have all the values higher than the corresponding values produced by the traditional methods mentioned above.

The predictors used in predicting the dependent variable in most of the research studies [1-9] include common variables such as population of the region, age group, income, among others; we are adding along with those variables, firearms rate i.e., number of gun ownership in every county, prison population and climate effect, to see if they have a significant impact on the crime rates. In order to corroborate our claim, we have studied [10-17] and concluded four of the six studies found the prevalence of firearms to be significantly and positively associated with homicide rates, and these associations were found across reasonably independent data sets. We observed a robust correlation between higher levels of gun ownership and higher firearm homicide rates. As for effect of climate on crime rates, [18-22] shows that weather has a strong causal effect on the incidence of criminal activity. With the help of Poisson Regression approach, the author [18] documented a striking relationship between weather patterns and crime rates and concluded higher temperature causes more crime. However, the study doesn't include any other parameters. Janet Gamble [23] studied the relationship between crime rates and weather in the city of Dallas using

numerable predictors and found that the relationship was not linear, but moderate and at high temperatures it becomes negative. Studies [24-26] found that assault is more common on weekends and holidays and found to be varying with years, months seasons and day length. Our research will consider all the factors mentioned and implement non-linear statistical model to understand which factors mostly influence the response.

3. DATA COLLECTION AND DESCRIPTION

We decided to use the data ranging from year 2008 to 2017 for 62 counties in the New York state. The following data were obtained from US census bureau and New York open data portal.

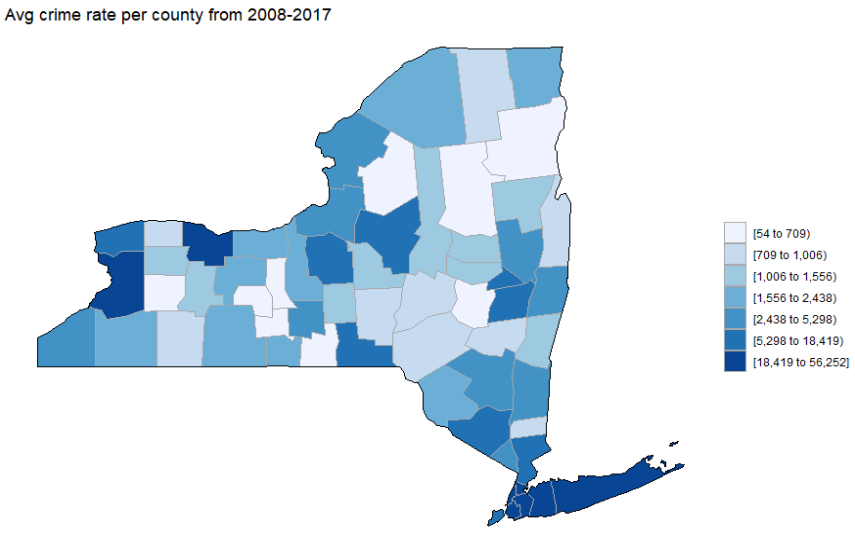
VARIABLE NAME	VARIABLE TYPE	VARIABLE DESCRIPTION
Crime.count.per.10000	Continuous	Response variable- The number of crimes per 10,000 population
Crime.type	Categorical	The crime type (9 Categories in total)
Average.of.Unemployment.Rate.per.10000	Continuous	The unemployment rate per 10,000 population
Population	Continuous	County population
prison.population.per.10000	Continuous	Prison population of a county per 10,000 population
firearm.count.per.10000	Continuous	firearm count per county per 10,000 population
firearm.rate.per.10000	Continuous	firearm rate per county per 10,000 population
total.male.per.10000	Continuous	total male population per 10,000 in a county
total.female.per.10000	Continuous	total female population per 10,000 in a county
WA.female.per.10000	Continuous	Number of white female per 10,000 population
WA.male.per.10000	Continuous	Number of white male per 10,000 population
BA.female.per.10000	Continuous	Number of Black female per 10,000 population
BA.male.per.10000	Continuous	Number of Black female per 10,000 population
AA.female.per.10000	Continuous	Number of Asian alone female per 10,000 population
AA.male.per.10000	Continuous	Number of Asian alone male per 10,000 population
H.male.per.10000	Continuous	Number of Hispanic male per 10,000 population
H.female.per.10000	Continuous	Number of Hispanic female per 10,000 population
BAC.female.per.10000	Continuous	Black or African American alone or in combination female population per 10,000
BAC.male.per.10000	Continuous	Black or African American alone or in combination male population per 10,000
WAC.female.per.10000	Continuous	White alone or in combination female population per 10,000
WAC.male.per.10000	Continuous	White alone or in combination male population per 10,000

The variables were normalized as their counts per 10,000 populations so that it would be more robust and also standardized. The crime types were consolidated into a single categorical variable.

4. DATA VISUALIZATION

1. **Heat map showing the total crime rates of each county:**

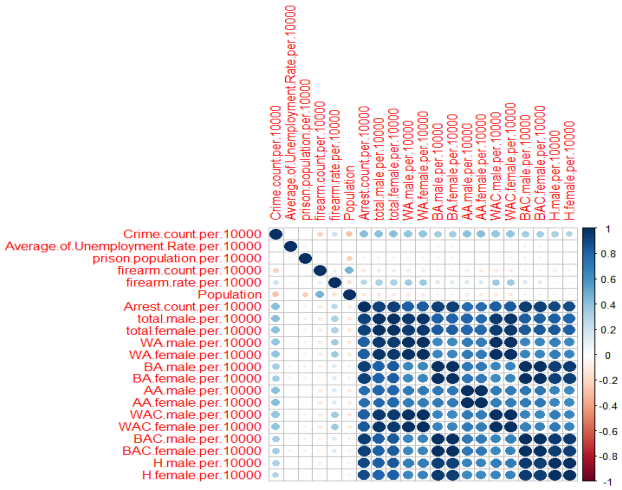
   This plot depicts the crime rates of the counties of New York State with varied opacity. The counties with higher opacity have higher average crime rates and the counties with lower opacities has a comparatively lower average crime rate. To visualize the crime rates of NY counties we used a choropleth map. From the choropleth map of New York State, it can be inferred that counties such as Erie, Monroe, Suffolk and Nassau has the highest average crime rate and Hamilton, Essex, Lewis and Wyoming has the lowest average crime rates.



Heat map

2. **Correlation plot:**

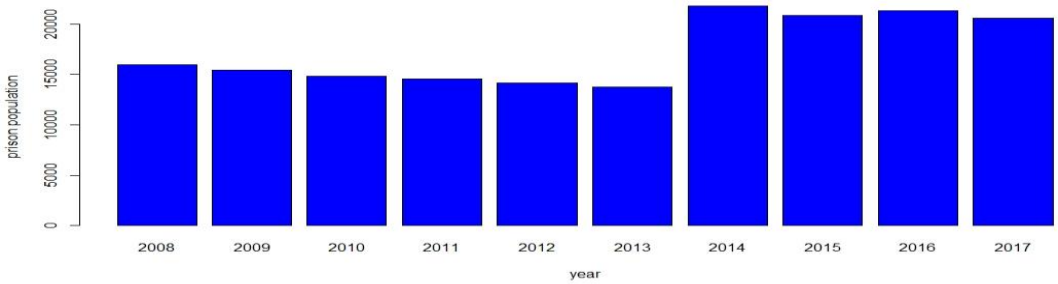
The correlation plot was done initially as part of our exploratory analysis to see if there any predictors are highly correlated with the response variable and also to see if the predictors are highly correlated among each other. Almost all of the predictors considered seem to have a positive correlation with the crime rate except for population and fire-arm count which seem to have a negative correlation. The demographic factors seem to have high correlation among each other. We decided to analyze both cases where we incorporate all of them in our model to see if the model is able to pick up any important variable and to remove the highly correlated variables to see if it makes a difference.



Correlation plot

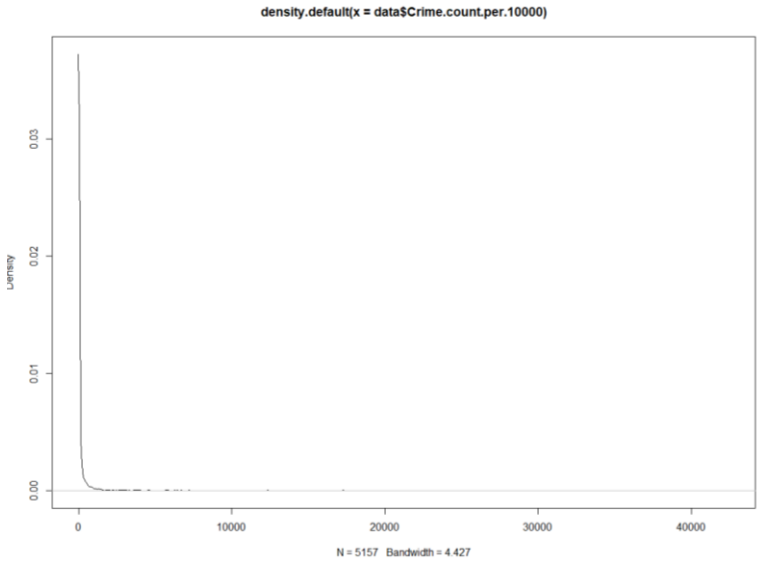
3. **Total prison population of the New York State:**

The bar graph of the total prison population of each year for 10 years from 2008 to 2017 is shown below. It shows that the prison population has been on a slight decrease till 2013 but had a sudden spike since 2014 and is sporadic from then with minimum changes.

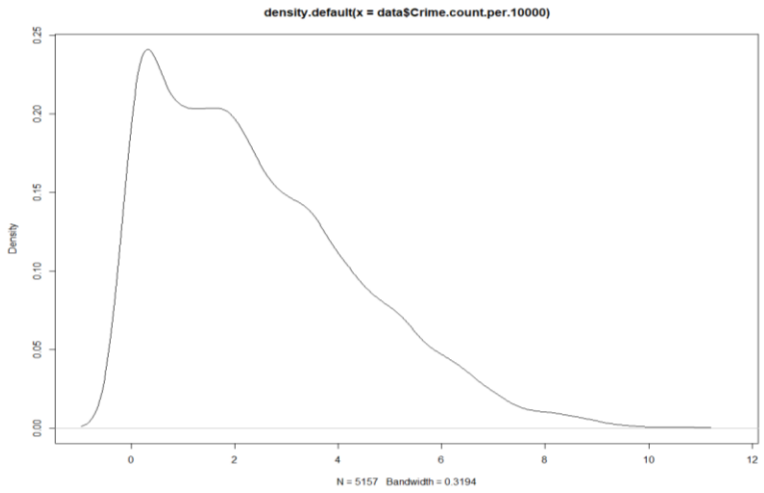


4. **Transformation of the response variable:**

   The density plot was first visualized for the crime rate per 10,000 population and it is as shown below.



   It is seen that the response variable follows a power log distribution and it is very hard to capture the effects of such a distribution. Hence a log transformation was done on the response variable and the resulting density plot of the response variable is as below.



The distribution after transformation is almost normal with skewness towards the right.

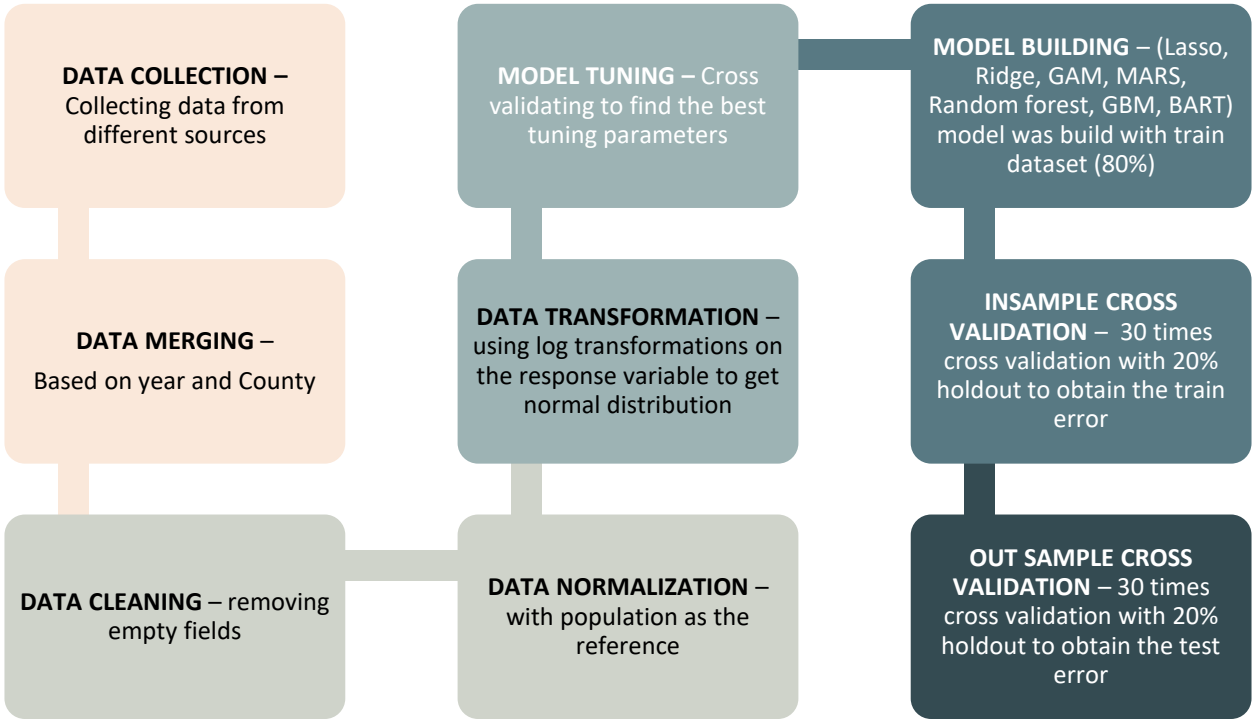
5. METHODOLOGY:

THE GENERAL APPROACH TO MODEL BUILDING:

The data has been collected from different sources, the socio-economic factors such as unemployment rate, jail population, adult arrests, fire-arms rates were obtained from the NY open data portal. The demographic data was obtained from US Census Bureau. These data were then consolidated and cleaned to form the final dataset. Every variable has been normalized as to per 10,000 populations to make the data more robust for prediction. A log transformation was applied on the response variable to obtain a convenient distribution for fitting the models. As the response variable is known we used supervised learning for prediction.

Both linear and nonlinear models were used to understand how well they can explain the variation in the data. Lasso, Ridge, GAM, MARS, Random forest, GBM and BART models were built. The parameters for every model were cross validated to find which resulted as the best values. The train data (80%) of the total number of observations is used for training each model 30 times to obtain the train error. At the same time the 20% holdouts are tested the same

number of times and test error is calculated as the mean of the 30 values obtained. The model with the lowest test error, which in our case GBM, is selected as the final model. Root Mean Squared Error and Mean Absolute error served as evaluators to decide on the best model.



The general approach to model building

SUPERVISED LEARNING:

We used supervised machine learning techniques in our project as our response variable which is the crime rate of counties in New York is known and labelled. Following are the models we implemented and ran diagnostics on each of them to understand which model performs better for the data that we have.

I. RIDGE REGRESSION:

Ridge regression is also a type of Linear regression which creates a reduced model when the number of predictor variables is greater than the number of observations or when there is multicollinearity. The ridge regression uses L_2 regularization which equals to the square of the magnitude of coefficients. Here the tuning parameter $\lambda = 0$. This does not eliminate any coefficients from the model and hence Lasso regression is easier to interpret than ridge. Ridge regression penalizes the size of the regression coefficients Specifically, the ridge regression estimate β_{hat} is defined as the value of β that minimizes

$$\sum_{i=1}^{n}(\text{yi} - \sum_{j}^{n}(xij - \beta_j)\text{^}2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

II. LASSO REGRESSION:

Lasso regression is a type of Linear regression with shrinkage parameter where the data points converge to a central point. It is a model with fewer parameters. It undergoes a L_1 regularization prior which takes a penalty of the absolute values of the magnitude of coefficients. The parsimonious model has fewer coefficients as the larger penalties will nullify the values closer to zero. Here the tuning parameter $\lambda = 1$

$$\sum_{i=1}^{n}(\text{yi} - \sum_{j}^{n}(xij - \beta_j)\text{^}2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

III. GENERALIZED ADDITIVE MODEL (GAM):

Generalized additive models (GAM) framework is based on a simple model:

1) Relationships between the individual predictors and the dependent variable follow smooth patterns that can be linear or nonlinear.

2) We can estimate these smooth relationships simultaneously and then predict g(E(Y))) by simply adding them up.

Mathematically speaking, GAM is an additive modelling technique where the impact of the predictive variables is captured through smooth functions which—depending on the underlying patterns in the data—can be nonlinear.Mathematically GAM can be represented as:

$$g(E(Y)) = \alpha + \beta_1(x_1) + \cdots + \beta_p(x_p),$$

where Y is the response, E(Y) denotes the expected value, and g(Y) denotes the link function that links the expected value to the predictor variables $x_1, \ldots, x_p$. The terms $\beta_1(x_1), \ldots\ldots, \beta_p(x_p)$,denote smooth, nonparametric functions.

There are three primary reasons for using GAM, which are mentioned below:

1. Interpretability

2. Flexibility/Automation

3. Regularization

1. Interpretability: When we have nonparametric regression model, the interpretation of single variable does not depend on the values of other variables in the model. For instance, if Y increases with an increase in X1, then all the variables are constant. Hence, by looking at the output of the model, we can understand and interpret the effects of variables on response.

In inclusion, GAM has the ability to control the smoothness of the predictor functions. We can avoid wiggly, non-important independent variables by simply adjusting the level of smoothness, which shows the relationship between response and features smoothly, even though the data at hand might suggest otherwise. This is important from a viewpoint of model interpretation.

2. Flexibility/Automation: GAM model captures nonlinear patterns that a classic model would fail to capture. When fitting a parametric regression model, nonlinear patterns are captured through polynomials, which leads to clumsy model. We don’t have this problem with GAM model.

Predictor functions are automatically derived during model estimation. We don't need to know beforehand what type of functions we will need.

3. Regularization: GAM model allows us to control smoothness of the predictor function to prevent overfitting. By controlling the wiggliness, we can directly tackle the bias/variance trade off. Say, we want to fit a model which predicts Y given 'x'. The equation of the model is given as:

$$Y = \beta\lambda(x) + \varepsilon$$

Where $\beta\lambda(x)$ is the smoothing function. The level of smoothness is determined by smoothing parameter denoted by λ. The higher the value of λ, the smoother the curve. Smoother curve has more bias and less variance.

GAM is useful for our project which is crime rate prediction to capture the effects of different predictors as the model is more flexible than a linear model.

IV. MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS):

Just like GAM, MARS is a nonparametric regression procedure. MARS constructs functional relationship between dependent and independent variables from a set of coefficients and basis function that are obtained from the regression data. The general MARS model equation is given as:

$$y = f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)=1$$

where y is the predictive function, which consists of an intercept parameter (β_0) and weighted sum of multiple basis function $h_m(X)$. MARS implementation is two steps procedure which is kept on applying until a desired model is found. In the first step, we build the model and increase its complexity by adding a basis function until a maximum level of complexity has been reached. In

the second step, we apply a pruning procedure which removes the basis functions from the model that contribute least to the overall goodness of fit.

MARS algorithm has gained popularity for finding predictive models for data mining problems where the predictor variables do not exhibit simple and/or monotone relationships to the dependent variable of interest which is the reason why we will use it in our project to see if the relationships can be captured by the model.

V. SINGLE REGRESSION TREE:

These involve stratifying or segmenting the predictor space into a number of simple regions. The predictors and cut points are chosen in such a way that minimum RSS is obtained. Tree-based methods are simple and useful for interpretation.

It follows a top-down greedy approach which is also called as recursive binary splitting as each split results in two child nodes. The approach is top-down because it begins at the top of the tree and successively splits further down the tree. It is greedy because the best split is made at every particular step.

$$R_1(j,s)=\{X|Xj <s \} \text{ and } R_2(j,s)=\{X|Xj \geq s\}$$

Where R_1, R_2 are the regions after a binary split is made and S is the cut point.

The process is repeated in a way that the cut points and the predictors are selected to get minimum RSS. The tree is grown until a stopping criterion is reached. Finally, after the regions have been created, we predict the response for a given test observation using the mean of the training observations in the region to which that test observations belong. The fully-grown tree often causes over-fitting. To avoid this, we tend to prune the tree by removing the weaker links by cross validation to reduce the variance produced. A shallow tree with fewer splits might lead to lower

variance and better interpretation at the cost of increase in bias. The usage of single tree produces overfitting and results in large out of sample errors.

In our project accuracy of prediction is very important and hence we will try to use more sophisticated methods such as bagging, boosting, Random Forest to overcome this problem but it will come at the cost of loss of some interpretation.

VI. ENSEMBLE TREES:

BAGGING:

As the usage of a single decision tree results in higher variance, Bootstrap aggregation or Bagging is introduced. To reduce the variance, we generate a number of bootstrapped training data to grow multiple trees and average the results to get better results.

In the bagged model each tree utilizes 2/3rd of the observations and 1/3rd of the observations are left unused. These are called the out of bag observations and each of the observations are predicted with the trained model and averaged. The errors produced by these observations are called the out of bag error.

In our project the usage of bagging model will help us determine the important predictors that affect the crime rates to a great extent. The disadvantage to this model is, when there is a higher correlation among the predictors there is a higher chance while training that each tree would pick up the same characteristics because all the predictors are tried at every split.

RANDOM FOREST:

Random forest is a special case of bagging where the number of predictors chosen(M) at every split is less than the total number of predictors (P). The typical values of number of predictors(M) is 1/3rd or the square root of the total number of predictors(P). Random forest is used to rectify

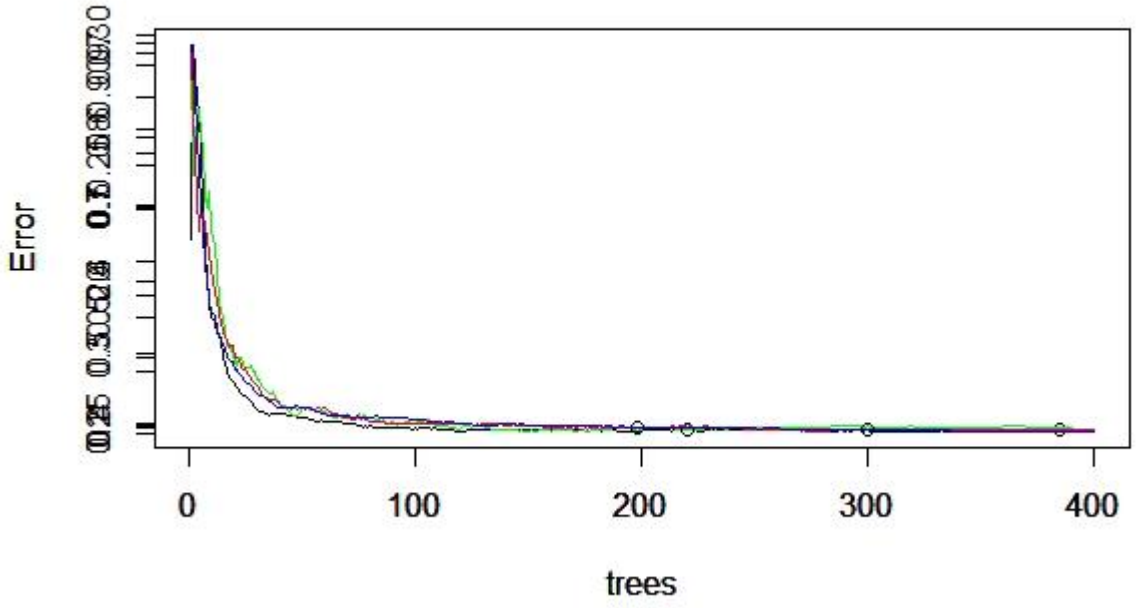
the disadvantages of Bagging. As we select the number of predictors(M) less than (P) we de-correlate the trees thereby reducing the variance.

STEPS INVOLVED IN RANDOM FOREST:

For a given data with 'N' observations and 'M' predictors random bootstrapped samples are created. 't' Deep regression Trees are grown using the optimal predictor among 'm' predictors for every split where m << M.

The predictor value is resulted from the average of the values obtained from 't' regression trees.

We will use Random forest in our project to see if it could produce any better results given the fact that it de-correlates the trees.



The random forest model has been built for mtry (number predictors considered during each split) = 4, 5, 7 and 21 as the total number of predictors we have is 21 and for the square root of 21 we

used both 4 and 5. One third of 21 is 7 so, we also tried mtry = 7. When cross validated the best fit was found for mtry = 7.

BOOSTING:

The problem with bagging models is that they tend to over fit as they grow bigger trees, alternatively boosting is used to reduce the over fitting by growing shallow trees. Given the model, it fits the decision tree to the residuals from the model. These decision trees are added sequentially and thus the fitted function is updated. It is a slow learner and gives more accurate results as it tries to reduce the error sequentially. There are several types of boosting models but we will use BART (Bayesian Additive Regression Trees) for our project.

Gradient boosting:

Gradient boosting is used for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

$$\text{Loss} = \text{MSE} = \sum(y_i - y_i^p)^2$$

Where y_i= ith target value, y_i^p = ith prediction, L(y_i, y_i^p) is Loss Function

By using gradient descent and updating our predictions based on a learning rate we minimize our loss function (MSE).

$$y_i^p = y_i^p + \alpha * \delta \sum(y_i - y_i^p)^2 \,/\, \delta y_i^p$$

Which becomes

$$y_i^p = y_i^p - \alpha * 2 * \sum(y_i - y_i^p)$$

Where, α is learning rate and $\sum(y_i - y_i^p)$ is sum of residuals.

So, the intuition behind gradient boosting algorithm is to repetitively leverage the patterns in residuals and strengthen a model with weak predictions and make it better. Once we reach a stage that residuals do not have any pattern that could be modelled, we can stop modelling residuals (otherwise it might lead to overfitting). Algorithmically, we are minimizing our loss function, such that test loss reach its minima.

BAYESIAN ADDITIVE REGRESSION TREES (BART):

BART is another type of model which incorporates boosting technique. It has two components, Sum of trees model, Regularization prior.

SUM OF TREES MODEL:

Sum of trees model can be mathematically represented as:

$$Y = \sum_{i=1}^{m} g_i(x_i T_i M_i) + \mathcal{E};\ \mathcal{E} \sim N(0, \sigma^2)$$

Where $M = \{\mu1, \mu2 \ldots \mu B\}$ denotes a set of parameter values

T denotes a binary tree

G(x,T,M) denotes a function corresponding to (T,M)

When the number of trees is large, the sum of trees model becomes more flexible and when coupled with regularization prior, the predictive accuracy increases dramatically.

REGULARIZATION PRIOR:

Regularization priors are used to control the model's complexity and to ensure that the final prediction does not depend upon a single tree. This eliminates the influence of a single tree on the prediction. Here T_i and $M_{i,b}$ is assumed to iid (independent and identically distributed). As this assumption is made, for every given tree T, a single μ and a single σ has to be chosen.

The tree prior is given by $\alpha(1+d)^{-\beta}$ where d corresponds to the depth of the node.

For the prior on μ, we start by shifting and scaling Y so that the probability E(Y|x) ϵ (-0.5,0.5). The value of σ is chosen such that 0.5 is within k standard deviations of zero. The prior increases the shrinkage of $\mu_{i,b}$ towards zero as m increases.

For the prior on σ, it follows an inverted chi-squared distribution: $\sigma^2 \sim \gamma\lambda/X_\gamma^2$. The usual value of γ ranges from 3 to 10 and for λ, a value 'q' has to be picked such as 0.75, 0.9 or 0.99 and set λ so that the q^{th} quantile of the prior on σ is located at the estimated value of $\hat{\alpha}$. Combining the prior distributions with the likelihood of the tree models, yields a posterior distribution. MCMC algorithm is used to characterize the posterior probability space.

The usage of BART in our project will help in obtaining a better accuracy of crime rate prediction as BART models are known to be more robust and better at prediction.

6. RESULTS:

We implemented linear and non-linear models to predict the number of crime rates based on the variables mentioned earlier and understand the relation between those. The best models were then cross validated for 30 times over a 20% holdout validation set. RMSE and MAE values of the validation set were used as the evaluators to determine the best model.

	RMSE		MAE		R-Squared	Percentage improvement compared to NULL (RMSE)	
	Train	Test	Train	Test		Train	Test
Ridge regression	2.522	1.135	1.967	1.135	0.672	-28.673469	42.09184
LASSO regression	2.525	1.333	1.969	0.879	0.681	-28.826531	31.9898
GAM	0.393	1.056	0.293	0.353	0.957	79.9489796	46.12245
MARS	0.462	0.478	0.346	0.351	0.95	76.4285714	75.61224
Random Forest	0.243	0.444	0.243	0.444	0.947	87.6020408	77.34694
GBM	0.238	0.325	0.182	0.241	0.985	87.8571429	83.41837
BART	0.293	0.35	2.16	0.261	0.972	85.0510204	82.14286
NULL	1.96	1.964	1.59	3.86			

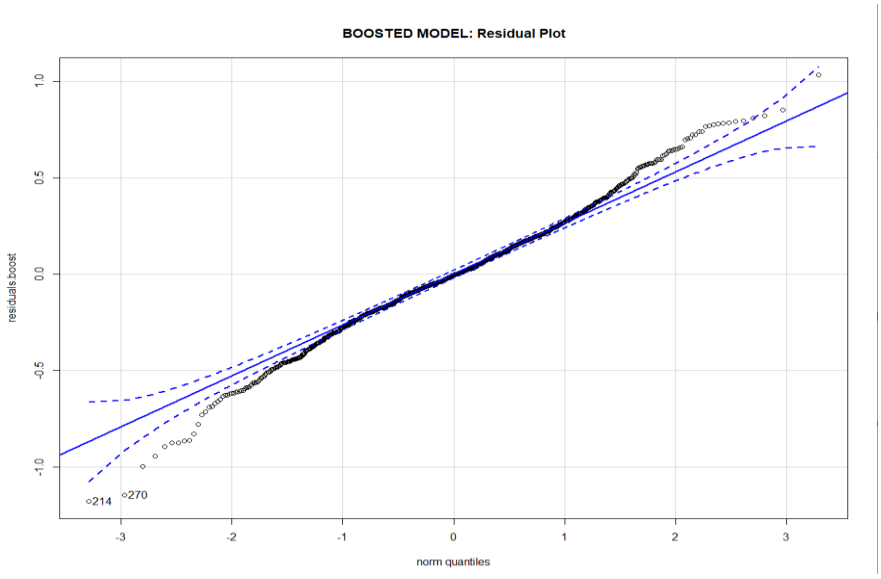
The table above shows the result of all the models we implemented. The best model proved to be Gradient Boosting Method based on low RMSE, low MAE and high R-squared value. It can also be noted that BART performs equally well but since for our research, inference is given more importance than prediction, we decided to choose GBM as the best model.

Variable Importance*:*

```
There were 21 predictors of which 21 had non-zero influence.
> summary(gbm1)
                                                                 var    rel.inf
Crime.type                                                 Crime.type 42.1495700
Arrest.count.per.10000                         Arrest.count.per.10000 34.3149002
total.female.per.10000                         total.female.per.10000  6.8265711
total.male.per.10000                             total.male.per.10000  4.7627112
BA.female.per.10000                               BA.female.per.10000  1.8404378
BAC.female.per.10000                             BAC.female.per.10000  1.6704088
WA.male.per.10000                                   WA.male.per.10000  1.6138686
BA.male.per.10000                                   BA.male.per.10000  1.1892309
WA.female.per.10000                               WA.female.per.10000  1.0091334
Population                                                 Population  0.7685306
WAC.female.per.10000                             WAC.female.per.10000  0.7207369
BAC.male.per.10000                                 BAC.male.per.10000  0.6683562
Average.of.Unemployment.Rate.per.10000 Average.of.Unemployment.Rate.per.10000  0.4795958
WAC.male.per.10000                                 WAC.male.per.10000  0.4005786
H.male.per.10000                                     H.male.per.10000  0.3077869
prison.population.per.10000               prison.population.per.10000  0.2466603
AA.male.per.10000                                   AA.male.per.10000  0.2428466
firearm.rate.per.10000                         firearm.rate.per.10000  0.2422291
firearm.count.per.10000                       firearm.count.per.10000  0.2096646
H.female.per.10000                                 H.female.per.10000  0.1876849
AA.female.per.10000                               AA.female.per.10000  0.1484976
>
```

Variable importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the variable. The three most important variable in our model are Crime type, Arrest.count.per.10000 and total.female.per.10000.

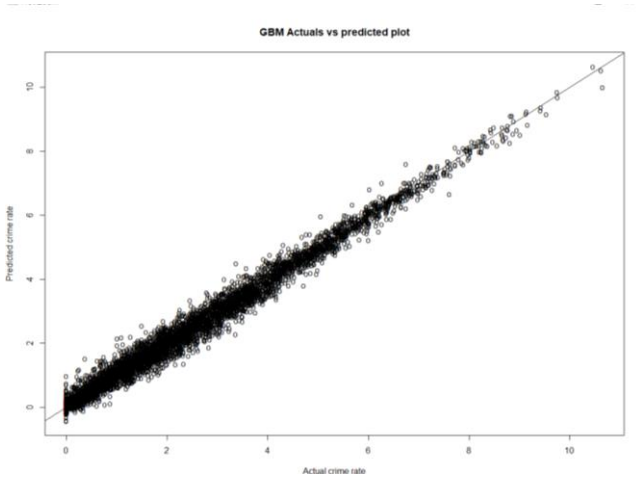
Residual Plot:



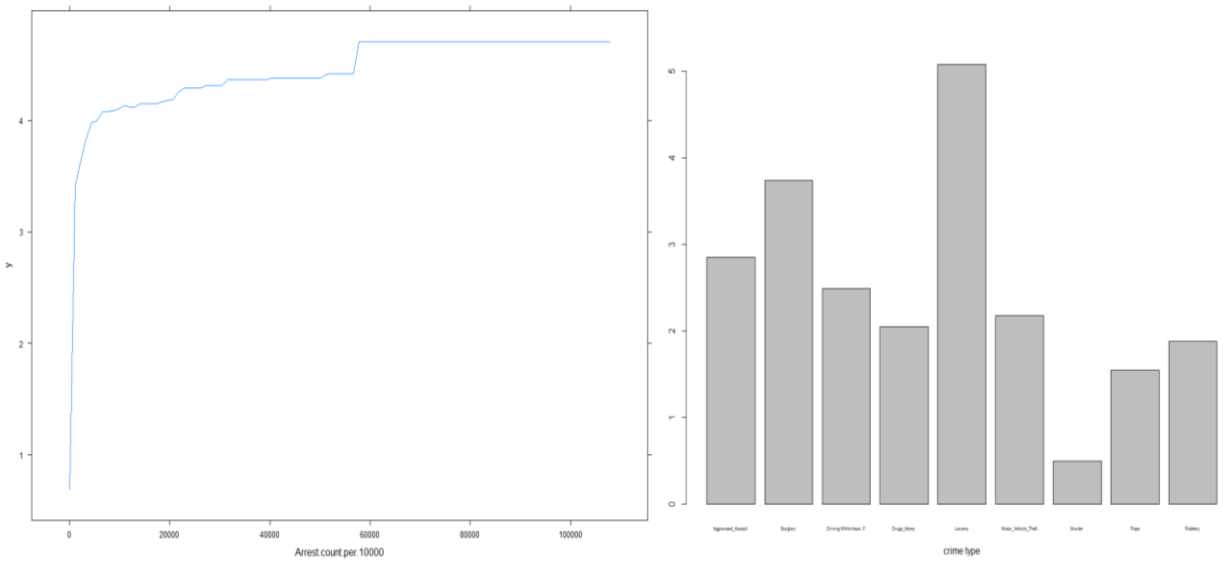
The residual plot of Gradient boosting model shows that most of the residuals fall with the 95% exhibiting nothing unusual and that the model is able to explain the variance.

Actual vs Predicted Plot:

The actual versus predicted plot demonstrates the values exhibiting the linear trend, which explains that the gradient boosting model does a good job of predicting the response as the points are tightly packed around the line.



Partial dependence plots:



The partial plots for the first two most importable predictors were drawn. It can be seen that the Larceny type has the highest influence among the other categories of the crime type. Also, for the arrest counts, there is a sharp increase in the crime rate when the arrest rate goes above 60,000. The rest of the partial dependence plots do not yield any significant conclusions.

7. CONCLUSIONS AND DISCUSSIONS:

On comparison with the other models, Gradient Boosting Model is found to have the lowest error and hence is chosen as the best model. The predictors that influence the crime rate prediction to a large extent are the crime type and adult arrests; whereas, jail population and fire-arms are not found to have a significant influence on the crime rate. This project can be used as a framework to predict the crime rates for different states or countries. As a future extension, other socio-economic factors such as household income, education level could be incorporated to enhance the crime rate prediction.

REFERENCES

[1] Van B. Shaw, Relationship Between Crime Rates and Certain Population Characteristics in Minnesota Counties, 40 J. Crim. L. & Criminology 43 (1949-1950)

[2] Varvara Ingilevicha, Sergey Ivanovb, Crime rate prediction in the urban environment using social factors, Procedia Computer Science, Volume 136, 2018, Pages 472-478

[3] L.G.A. Alves, H.V. Ribeiro, F.A. Rodrigues, Crime prediction through urban metrics and statistical learning, Physica A, 505 (2018), pp. 435-443

[4] Rizwan Iqbal1, Masrah Azrifah, Azmi Murad, Aida Mustapha, Payam Hassany, Shariat Panahy, Nasim Khanahmadliravi, An Experimental Study of Classification Algorithms for Crime Prediction, Indian Journal of Science and Technology.

[5] Kumar V, and Rathee N (2011). Knowledge discovery from database using an integration of clustering and classification, International Journal of Advanced Computer Science and Applications, vol 2(3), 29–32.

[6] Kang H-W, Kang H-B (2017) Prediction of crime occurrence from multi-modal data using deep learning. PLoS ONE 12(4): e0176244.

[7] Chen P, Yuan H, Shu X. Forecasting Crime Using the ARIMA Model. In: Proceedings of the 5th IEEE International Conference on Fuzzy Systems and Knowledge Discovery. vol. 5; 2008. p. 627–630.

[8] Liao R, Wang X, Li L, Qin Z. A novel serial crime prediction model based on Bayesian learning theory. In: Proceedings of the 2010 IEEE International Conference on Machine Learning and Cybernetics. vol. 4; 2010. p. 1757–1762.

[9] R. Liao, X. Wang, L. Li, and Z. Qin, "A novel serial crime prediction model based on Bayesian learning theory," 2010 Int. Conf. Mach. Learn. Cybern., no. July, pp. 1757–1762, 2010.

[10] The Relationship Between Firearm Prevalence and Violent Crime https://www.rand.org/research/gun-policy/analysis/supplementary/firearm-prevalence-violent-crime.html

[11] Bice, Douglas C., and David D. Hemley, "The Market for New Handguns: An Empirical Investigation," Journal of Law & Economics, Vol. 45, No. 1, 2002, pp. 251–265.

[12] Kleck, G., and E. B. Patterson, "The Impact of Gun Control and Gun Ownership Levels on Violence Rates," Journal of Quantitative Criminology, Vol. 9, No. 3, 1993, pp. 249–287.

[13] Kleck, Gary, Tomislav Kovandzic, and Mark E. Schaffer, Gun Prevalence, Homicide Rates and Causality: A GMM Approach to Endogeneity Bias, London: Centre for Economic Policy Research, Discussion Paper No. 5357, 2005.

[14] The relationship between gun ownership and firearm homicide rates in the United States, 1981-2010. https://www.ncbi.nlm.nih.gov/pubmed/24028252

[15] Ronald J. Frandsen, Dave Naglich, Gene A. Lauver, Regional Justice Information Service, Allina D. Lee, Bureau of Justice Statistics.

[16] https://www.bjs.gov/ucrdata/Search/Crime/Crime.cfm

[17] Siegel, M., C. S. Ross, and C. King, "Examining the Relationship Between the Prevalence of Guns and Homicide Rates in the USA Using a New and Improved State-Level Gun Ownership Proxy," Injury Prevention, Vol. 20, No. 6, 2014, pp. 424–426.

[18] Matthew Ranson, Crime, Weather, and Climate Change, Journal of Environmental Economics and Management, Volume 67, Issue 3, May 2014, Pages 274-302.

[19] Chris Brunsdon, Jonathan Corcoran, Gary Higgs, Andrew Ware,
The influence of weather on local geographical patterns of police calls for service
Environ. Plan. B: Plan. Des., 36 (5) (2009), pp. 906-926

[20] Brad Bushman, Morgan Wang, Craig Anderson

Is the curve relating temperature to aggression linear or curvilinear?

J. Personal. Soc. Psychol., 89 (1) (2005), pp. 62-66

[21] Ellen Cohn, Weather and crime, Br. J. Criminol., 30 (1) (1990), pp. 51-64

[22] James Horrocks, Andrea Menclova, The effects of weather on crime, N. Z. Econ. Pap., 45 (3) (2011), pp. 231-254

[23] Gamble JL, Hess JJ. Temperature and violent crime in dallas, Texas: relationships and implications of climate change. West J Emerg Med. 2012;13(3):239-46.

[24] Cohn EG, Rotton J. Assault as a function of time and temperature: A moderator variable time-series analysis. J Pers Soc Psychol. 1997;72:1322–34.

[25] Rotton J, Cohn EG. Violence is a curvilinear function of temperature in Dallas: A replication. J Pers Soc Psycho. 2000;l78:1074–81.

[26] Rotton J, Cohn EG. Temperature, routine activities, and domestic violence: A reanalysis. Violence and Victims. 2001;16:203–15.

[27] https://businessjargons.com/methods-of-determining-correlation.html