CRIME RATE PREDICTION OF COUNTIES OF NY STATE USING PREDICTIVE MODELING

Team 8:
Aiyngaran Chockalingam
Chandan Singh
GuruPraneeth Rao
Sandeep Purushotham

University at Buffalo The State University of New York

Agenda

- INTRODUCTION
- LITERATURE REVIEW
- DATA COLLECTION
- DATA TRANSFORMATIONS
- DATA VISUALIZATIONS
- METHODOLOGY USED
- MODELING RESULTS
- EVALUATING THE BEST MODEL
- CONCLUSION AND FUTURE SCOPE

INTRODUCTION

- The primary goal of the project is to predict the crime rates of counties of NY State and to determine the key factors that contribute to it.
- Incorporating data analytic techniques, it is possible to predict the crime rates and take necessary steps in curbing them. It is also necessary to understand the socio-economic and geographical contexts of a county to implement judicial decisions.
- Our analysis is implemented using county-level data on crime rates, population and fire arms rates, prison admissions, socio-demographics, and adult arrests during the period 2008-2017 for the state of New York

LITERATURE REVIEW

- An exhaustive literature review was conducted. It was noted that several socio-economic factors such as median household income, population, unemployment have shown to be positively correlated.
- Literature reviews on crime rate prediction using demographic data have shown that few races are seen to be significant predictors of crime rate.
- Unlike other previous literatures, we wanted to understand whether the number of arrests or the jail population in a county would have any co-relation with the crime rates.
- Also, another novelty of this project is the use of fire-arms data and to determine its influence on the crime rates.
- As these data were only available on a yearly basis, we could not incorporate weather data to determine the effect of seasonality.

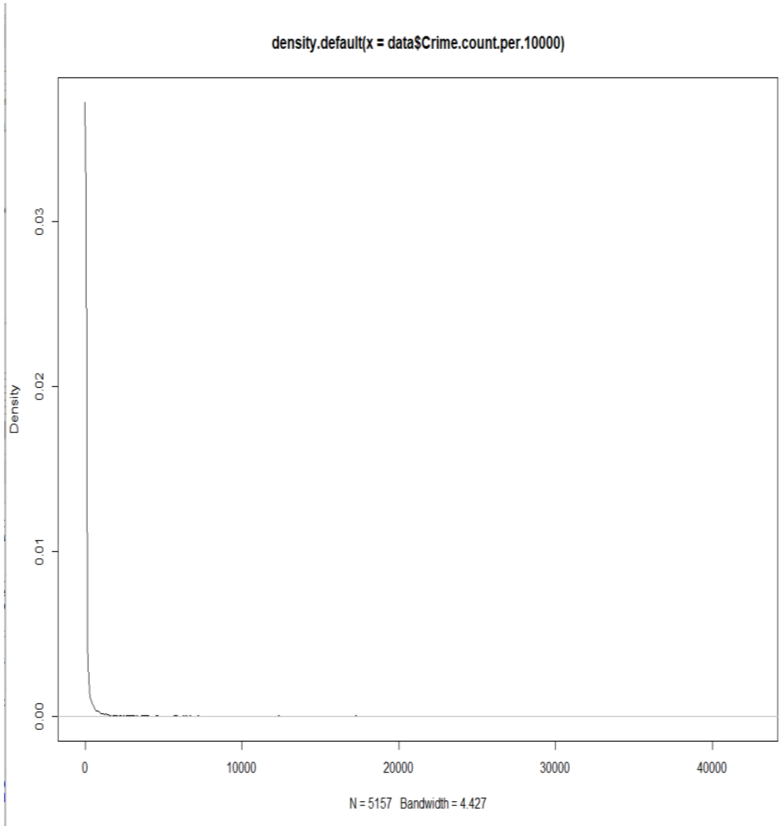
DATA COLLECTION

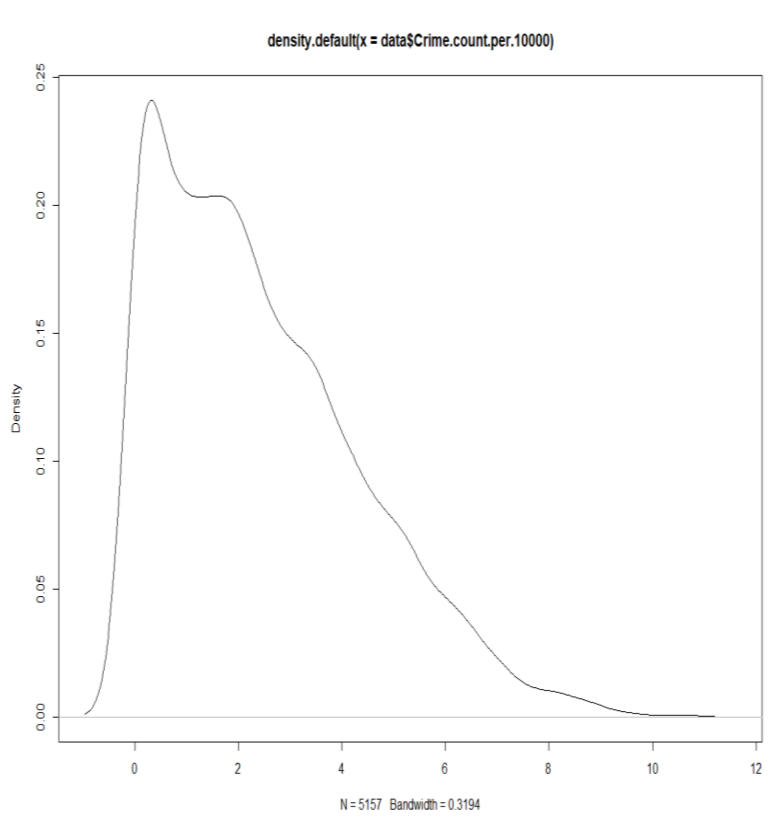
- Five different datasets were combined from different sources to obtain the final dataset that would answer the research question.
- County-level data for NY State for the years 2008-2017 was considered for this project.
- The socio-economic factors such as unemployment rate, jail population, adult arrests, fire-arms rates were obtained from the NY open data portal.
- The demographic data was obtained from US Census Bureau.
- These were consolidated to form the final dataset.

DATA TRANSFORMATIONS

- The crime rates were normalized as the number of crimes per 10,000 population as this was found to be a more robust method for prediction.
- The same transformations were applied on other variables to establish a generalized framework.
- The NA values were removed from the dataset as they were very less in number.
- The response variable initially had a power log distribution. A log transformation was carried out to facilitate better prediction accuracy.

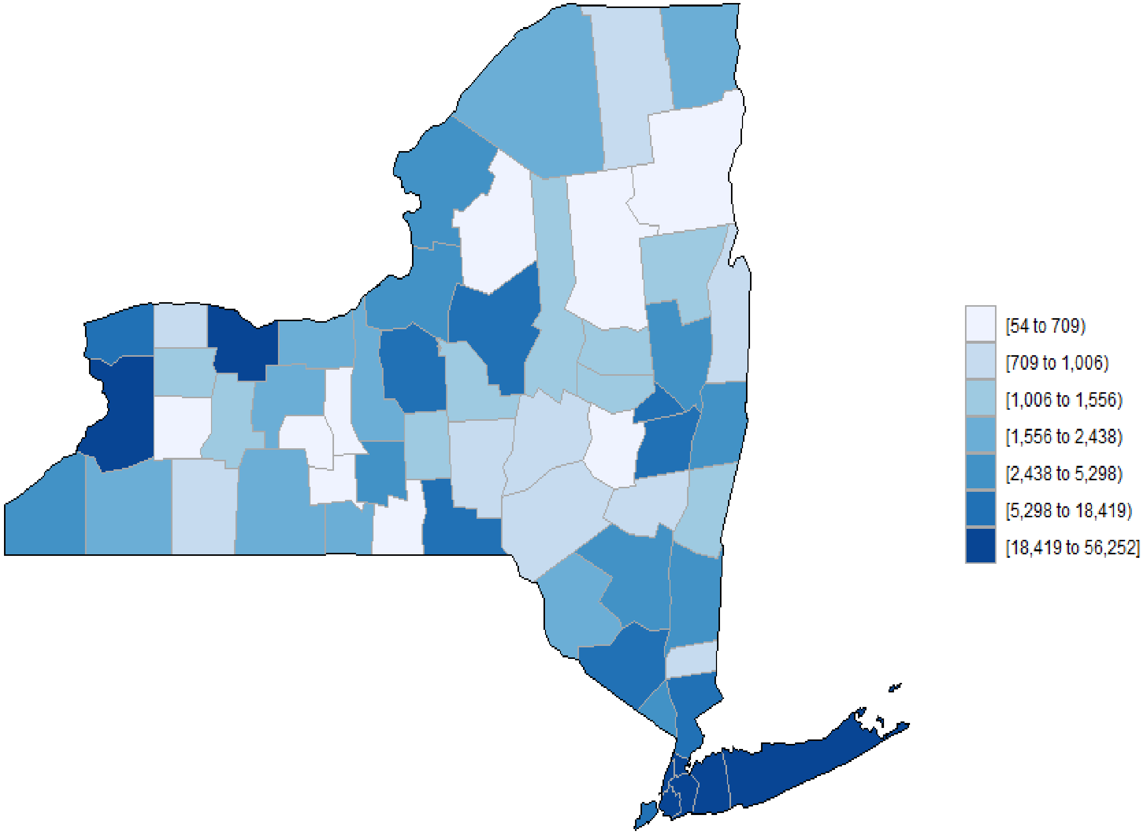
Response variable – Before and After



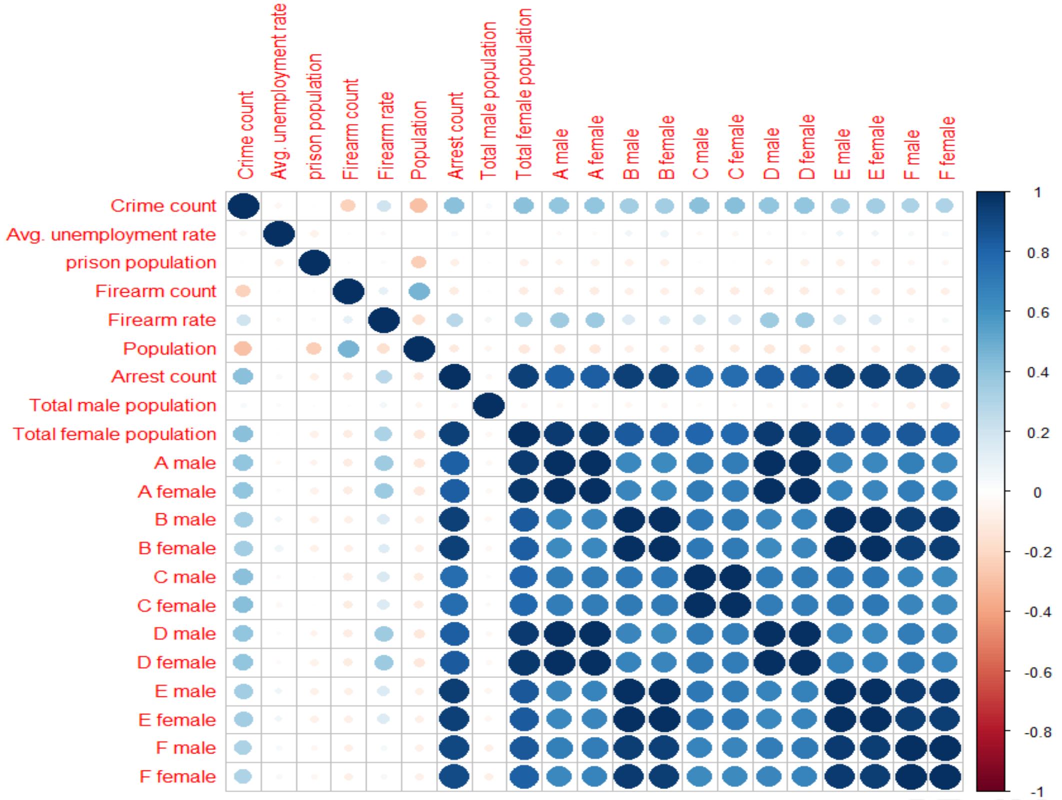


DATA VISUALIZATIONS

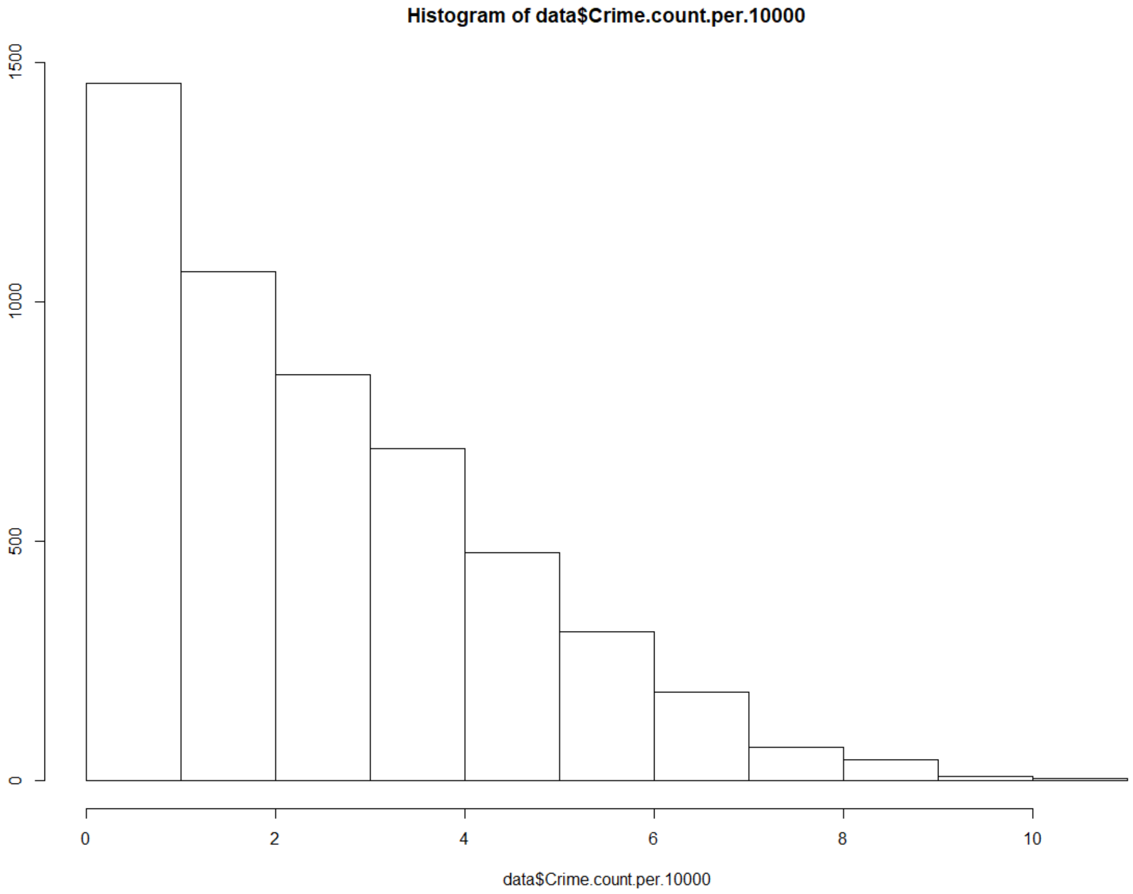
Average crime rate per county from 2008-2017



Co-relation plot



Histogram of the transformed response variable

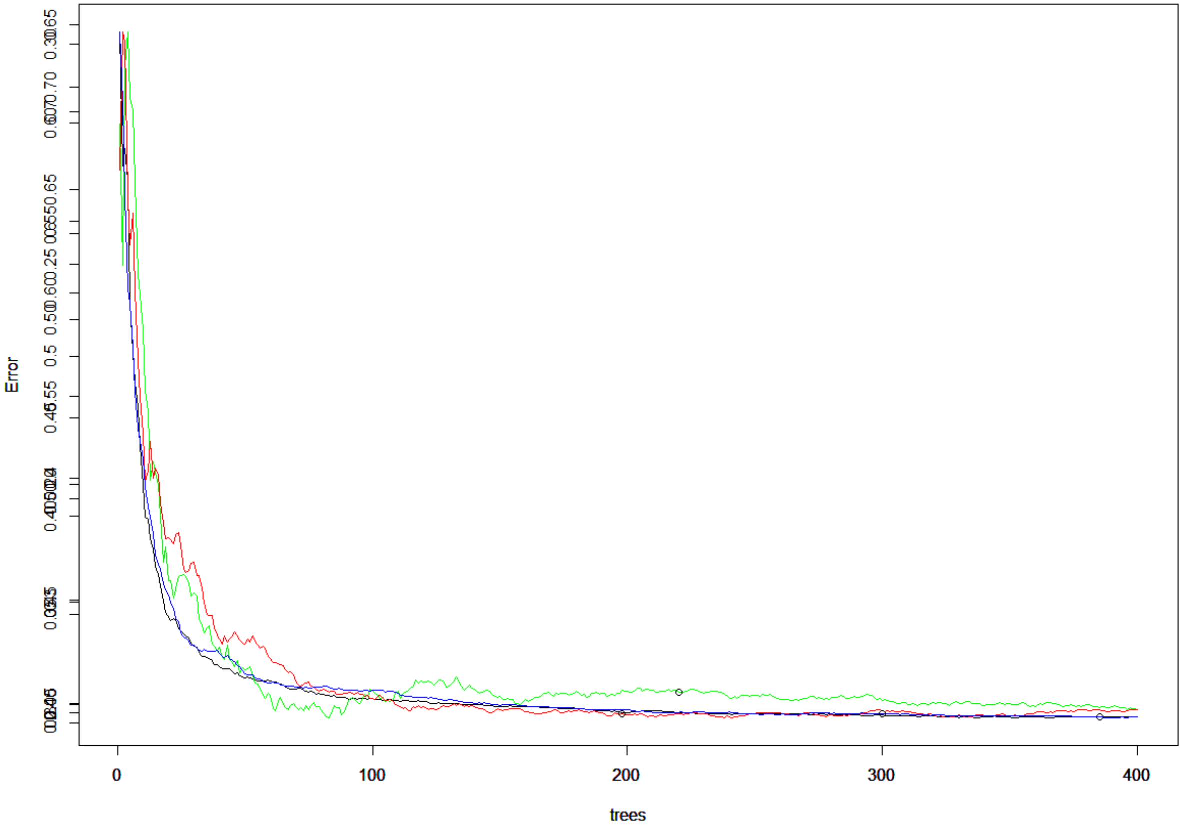


METHODOLOGY USED

- Both linear and non-linear models were used for the prediction to understand their behavior.
- The parameters of every model was tuned by Cross Validation techniques to obtain the best parameters.
- 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.
- Model diagnostics were run for each model to understand how each model was able to capture the variation in the data.

Comparison of different RF models

- Black line – mtry =7
- Red line – mtry = 5
- Green line – mtry = 4
- Blue line – mtry = 21

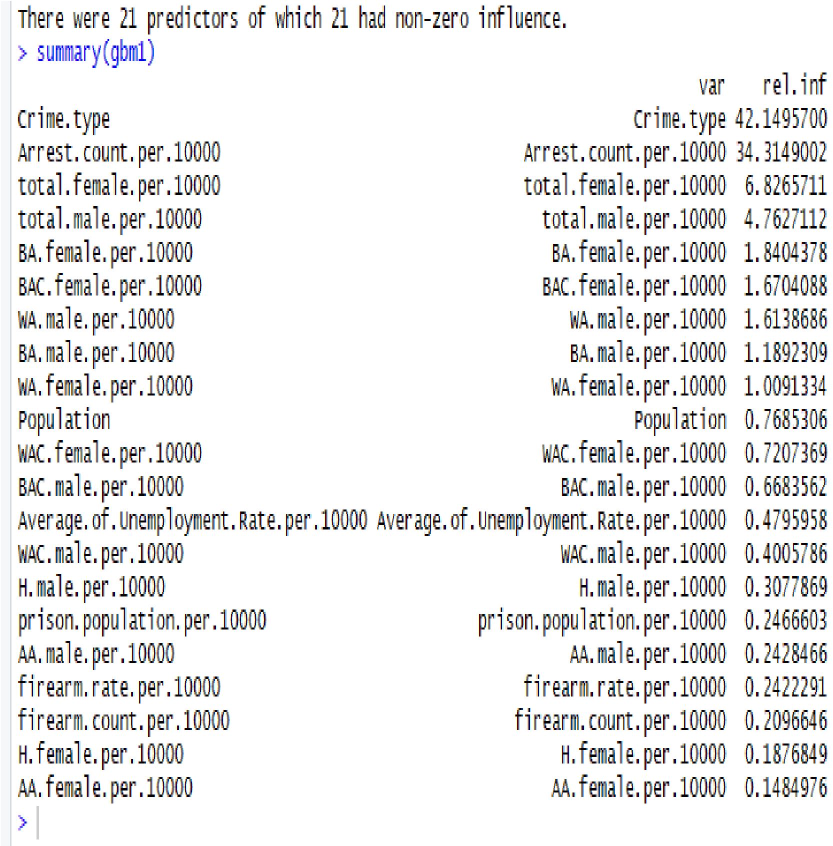


Model Results

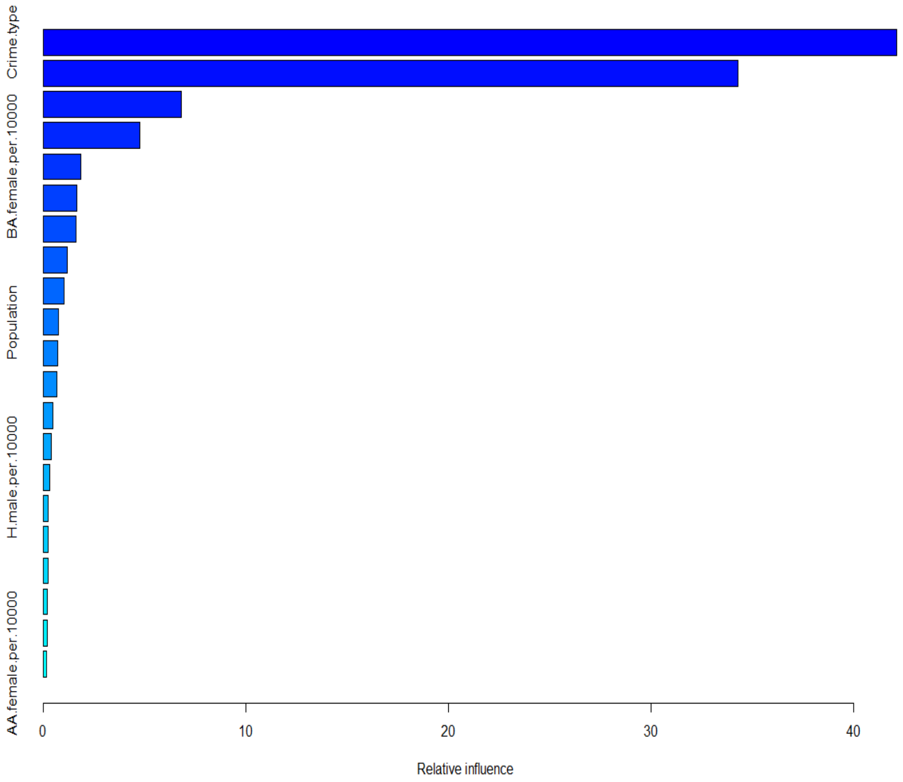
	RMSE		MAE		R-Squared	Percentage improvement compared to NULL (RMSE)	
	Train	Test	Train	Test		Train	Test
Ridge regression	2.522	1.135	1.967	1.135	0.672	-28.673469	42.09184
LASSO regression	2.525	1.333	1.969	0.879	0.681	-28.826531	31.9898
GAM	0.393	1.056	0.293	0.353	0.957	79.9489796	46.12245
MARS	0.462	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			

Model diagnostics for the best model

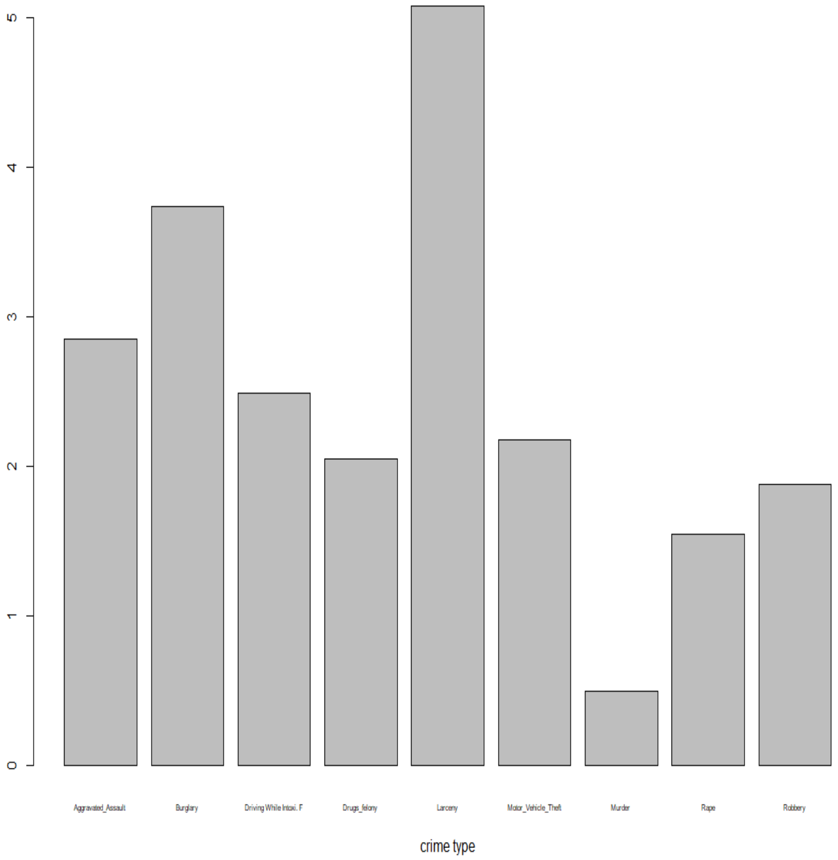
- The Best model chosen was GBM

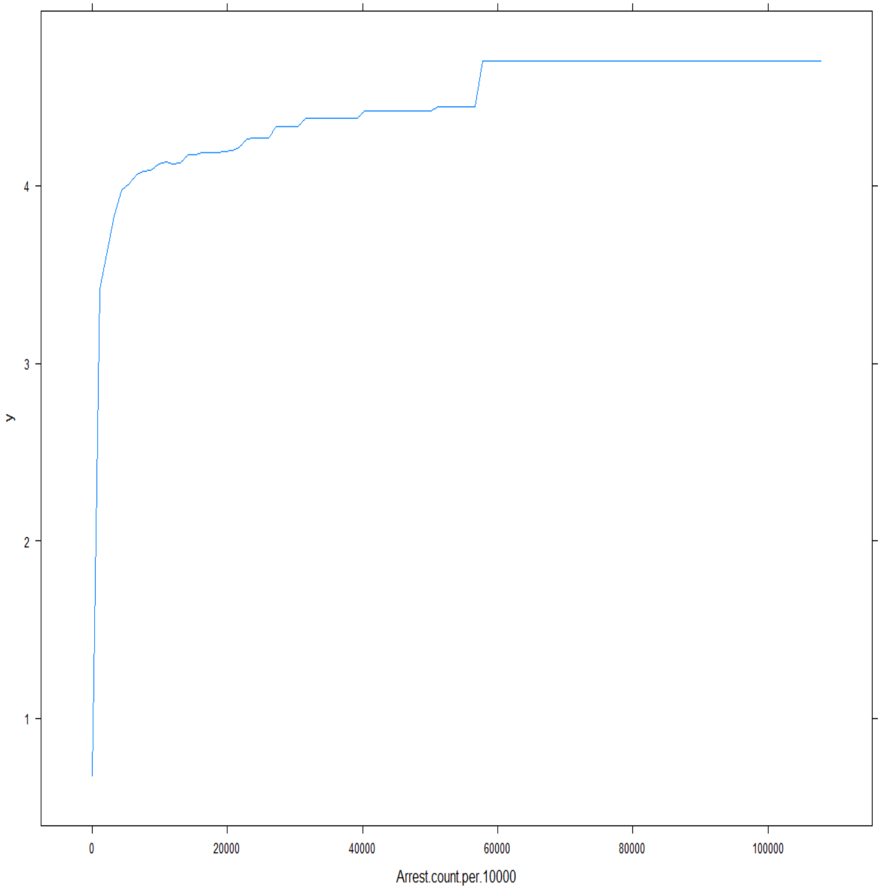


```
There were 21 predictors of which 21 had non-zero influence.
> summary(gbm1)
                                                                 var    rel.inf
Crime.type                                                Crime.type 42.1495700
Arrest.count.per.10000                        Arrest.count.per.10000 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
> 
```

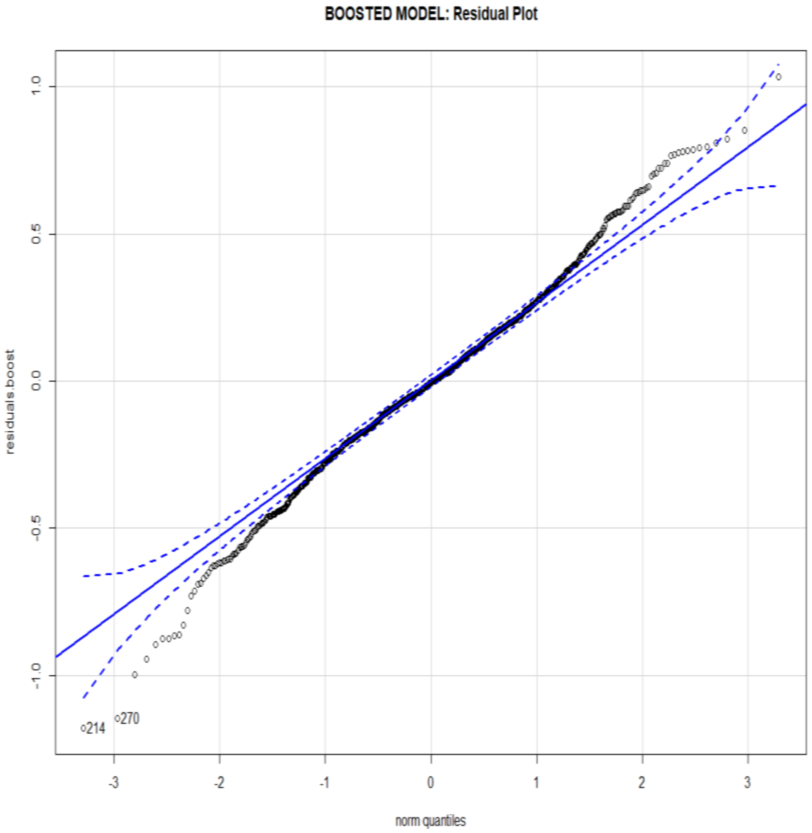


Partial dependence plots

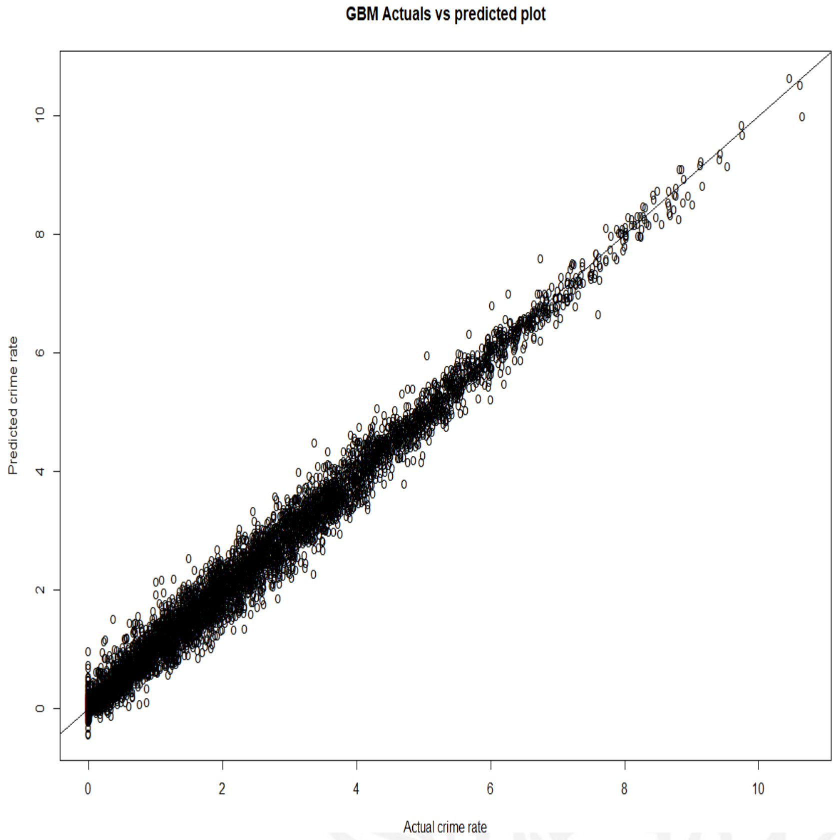




QQ Plot and Actuals vs predicted plot

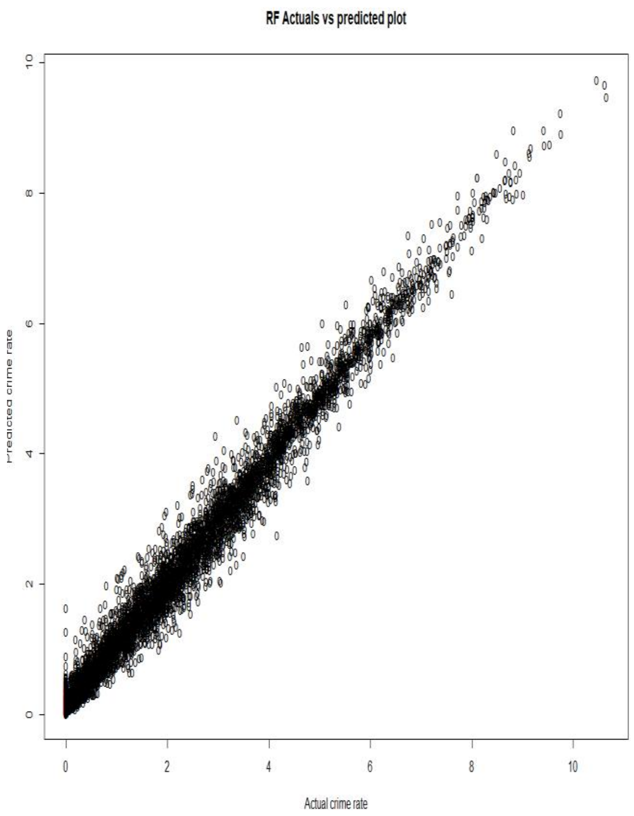


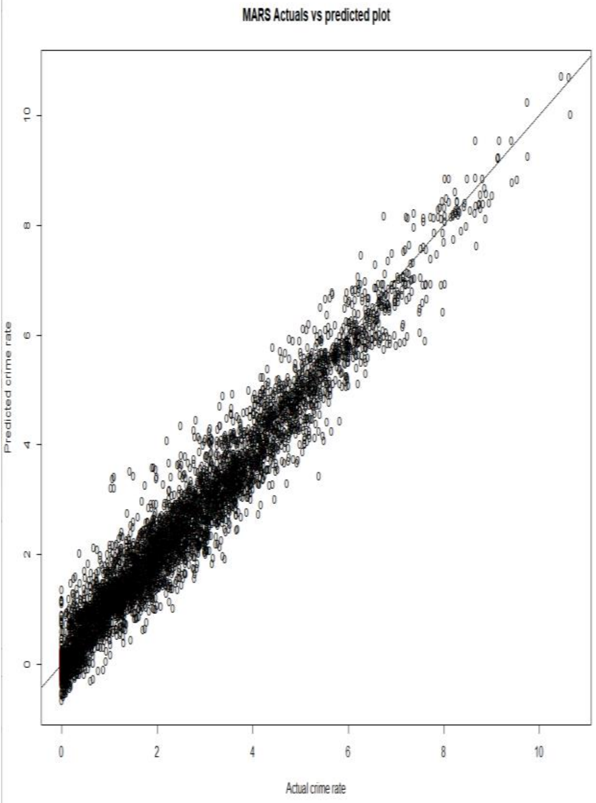
BOOSTED MODEL: Residual Plot

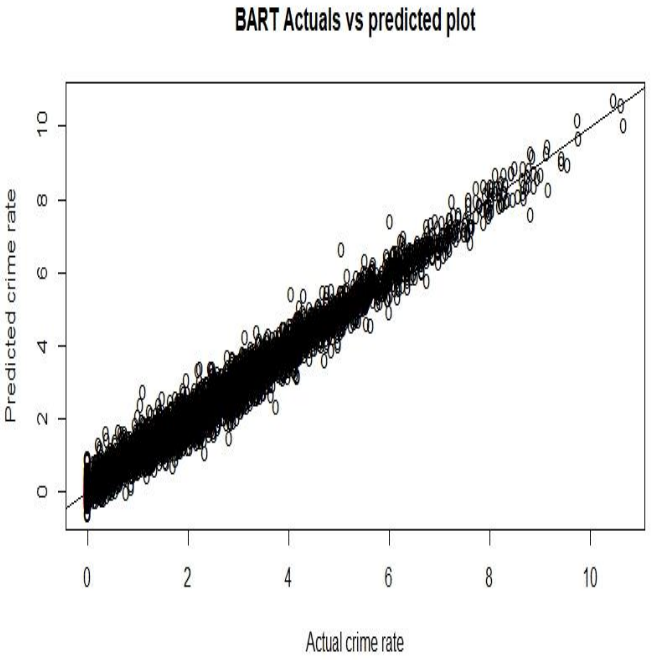


GBM Actuals vs predicted plot

Actuals vs predicted plots of other models







Conclusions and future scope

- On comparison with the other models, gbm 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.
- Jail population and fire-arms are not found to have a significant influence on the crime rate.
- This project can be established for different states. Other socio-economic factors such as household income, education level could be incorporated to enhance the crime rate prediction.