P8130 Final Project Report

Jinghan(jl6048), Paula (pw2551), Yuxuan(yc4018), Yuan(ym2866), Yunlin(yz4184)

Abstract

In this project, we evaluated how population characteristics are related to the crime rate across US counties and what are the possible predictors that contribute to the crime rate. To reach our goal, we performed data preprocessing and selected several candidate models for crime rate predicting. Our final model includes potential predictors including total population, youth concentration, percent of Bachelor's degrees and high school graduates, poverty and unemployment rate, per capita income, population density, region, hospital beds ratio, and interaction term of population and percent of Bachelor's degrees.

Introduction

Crime is one of the major problems in modern society as it brings disequilibrium to the normal life of the general population. Moreover, the rate of committing crime is related to several factors including economic conditions, population density, the youth concentration in the general population, etc [1]. Despite the large research literature addressing various aspects of these issues, we also want to dive into this common societal problem and give our conclusion. In this project, we are interested in using the "County Demographic Information" (CDI) data set to build a model that predicts the crime rate. In order to achieve our goal, we cleaned the data set and added useful variables. We later carried out several methods to compare different models and picked out the best one.

Methods

The dataset contained crime related statistics from 440 counties in the US collected between 1990-1992. We first cleaned the dataset and normalized some variables for better comparison. We added variables for crime rate, number of doctors, number of hospital beds, all per 1000 population. We also divided total population by land area to get population density and changed the variable "region" as dummy variables. We then removed variables that we have already performed transformation upon, which includes: ID, total number of crimes, land area, number of active physicians, number of hospital beds, the total personal income, and geographic region from our dataset. Our final dataset has 440 observations of 17 variables.

Later we plotted boxplots for each variable to examine the rough distribution of each numeric variable. We also plotted a scatterplot matrix and a correlation heatmap to visualize and check the marginal distribution and correlation among all variables. In the model-building process, we built and refined the full first, as it will be the basis of later analysis. We performed model diagnostics to check for influential points via Cook's distance plot and removed observation 1 and 6 as they were extremely deviated from the norm. We also found the λ, approximately equals to 0.5, to transform Y using the BoxCox transformation to improve the model fit.

Upon the newly-fitfull model, we used backward elimination, forward selection, and stepwise selection to explore useful predictors and build candidate regression models. We additionally fit three more models: the first by adding one interactive term of our choice to the backward elimination model, and the other two by using the criterion-based procedures (Adjusted R^2 and Mallows Cp value). For each of the models mentioned above, we used residual vs fitted/predicted values plot and QQ plot to check for violations of regression model assumptions; scale-location plot to check for the assumption of equal variance cook's distance plot to check for outliers; and multicollinearity to check the correlation between predictors. For a more straightforward presentation of model comparison, we calculated the adjusted R^2, AIC, BIC, Cp, and RMSE values for each of the models and presented them in a table. Last but not the least, we applied 10-fold cross validation to all models to evaluate our models' performance and plot the RMSE value distribution for each model.

Results

Before any visualization or statistical analysis, we first checked the missing values in the dataset shown in *Table 1*. There was no missing in this dataset, thus no data imputation needed. To give a first glance on our dataset, we arranged the state by the crime rate from low to high. As we can see in *Figure 1*, the lowest crime rate is in Pennsylvania and the highest crime rate is in Washington, DC. The mean of the crime rate is around 57 cases per 1000 population. When we grouped the data by county, we also found that Kings county of New York has an extremely high crime rate of nearly 296 cases per 1000 population. This data entry was proved as an outlier and was removed along with another entry in the later section.

Distribution of each variable

Before we divide further into establishing a regression model, let's first take a look at the distribution of each variable. As we can see in *Figure 2&3*, nearly every variable has outliers, and most variables are right skewed, especially in variables such as Population Density, Population, Crime Rate.

Marginal Correlation and Correlation Heatmap

We used the correlation matrix and correlation heatmap shown in *Figure 4&5* to plot the relationship between all variables to examine marginal distribution of each variable versus the crime rate and to confirm which variables are suitable for inclusion in our model. From the scatter plot matrix, we observe that there are strong correlations between our predictor variables. Thus, we took cautiousness in selecting variables and examined multicollinearity in each of the models generated in later steps.

Model Selection and Diagnostic

1. Full model before transformation

We built the full model and examined its model diagnostics first. Ideally, we would like to see that residual values bounce around 0 in the Residuals vs Fitted plot. However, the plot (*Figure 6*) showed that the fitting red line was bent and distorted, and that most of the dots are gathered on the left. There are also several extreme outliers that skewed the whole distribution, suggesting that we need some adjustions. We thus used the box-cox plot to get the lambda. From the plot (*Figure 7*), we can see that lambda roughly equals 0.576, thus we applied square root to the Y to adjust the linear model.

2. Full Model after transformation

After we applied the transformation to the full model, the residual values now bounce around 0 in the Residuals vs Fitted plot (*Figure 8*), which indicates that the linear assumption is met. There are no extreme outliers and the variance of the residuals is are equal. In model after transformation, the fitting line spreads equally along the range of predictors, our assumption of homoscedasticity is likely satisfied for a given regression model. The whole distribution has been adjusted.

3. Backward Elimination, Forward Selection, Stepwise Selection

We have used backward elimination, forward selection, and stepwise selection procedure to get three tentative models. We noticed that the model generated by forward selection is the same as the full model, while the model generated by using stepwise procedure is the same as the backward model. Thus, we will not continue discussing them from now on. We then plotted the model diagnostics (results similar to the full model's , *Figure 8*) to check whether this model met the linear assumption and its variances of error terms are equal. And according to the figure, the answer is yes. There are no outliers beyond the dashed lines.

4. Adding Interaction Term

We also wanted to examine whether there are interactions between population and percent of Bachelor's degrees. The choice of these two variables was mainly based on the correlation heatmap and avoiding high multicollinearity between variables at the same time. We added this term to the backward selected model. To show the term we added is significant, we conducted a partial F-test between this new model and the backward selected model. By getting a p-value of 0.005064 (*Table 2*), we rejected the null hypothesis and concluded that this interaction term is indeed significant. Similar diagnostics steps were applied and assumptions were all met (similar to the full model's , *Figure 8*).

5. Criterion Based Procedure: Models and Statistics

We selected the best model suggested by criterion-based procedures, which optimized some measures of goodness. We first compared the two plots of Cp and adjusted R^2 as functions of numbers of parameters (*Figure 9*). In this figure, we checked what is the best value of predictors by examining when Cp reached its lowest point and when adj. R^2 reached its apex. Based on *Figure 9*, Cp reached the smallest value when the number of predictors is 9, and the adj. R^2 reached the largest value when the number of predictors is 12. We thus fit two models according to the number of predictors and performed model diagnostics. Next, We extracted the adjusted R^2, Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallow's criterion value (Cp), and RMSE (*Table 3*) from all above-mentioned models and took them into consideration. Through this process, two nested candidate models were chosen: the model with interaction term has the largest adjusted R^2 with the smallest AIC, Cp, and RMSE; and the model which based on Cp value has the smallest BIC value.

6. Cross validation

To pick up the most efficient model among the 5 models, we applied 10-fold cross validation to all models to evaluate our models' performance and plot the RMSE value distribution for each model (*Figure 10*). Here we can see the interaction model has the smallest RMSE and thus we chose this model as our best model.

Conclusions/Discussion

In order to find the best model, we created several models by stepwise regression, adding interaction terms to the backward model with partial test to prove its significance, and criterion-based procedures. We then concluded that the interaction model is the best by comparing adjusted-R^2, AIC, BIC, Cp, and RMSE stepwise regression. Instead of just using one verification step, 10-fold cross-validation (CV) was applied to confirm the best model. Our final model contains potential predictors including population, youth concentration, percent of Bachelor's degrees and high school graduates, poverty and unemployment rate, per capita income, population density, region, hospital beds ratio, and interaction term of population and percent of Bachelor's degrees. The summary of the interaction model (*Table 3*) showed that while most predictors had positive correlations with crime rates, percent of Bachelor's degrees, northeast, northcentral, population and percent of Bachelor's degrees were negatively correlated with crime rate. As these predictors increase, the crime rate will decrease. Starting from the full model, we used several methods to diagnose and finally chose the interactive model as our best model. Nevertheless, there is a notable limitation. Even though the interactive model is the best fit model among those 5 models, the advantages of the interactive model are not that strong compared to other models.

Appendix

id	cty	state	area	pop	pop18	pop65	docs	beds	crimes	hsgrad	bagrad	poverty	unemp	pcincome	totalinc	region
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 1: Checking missing values

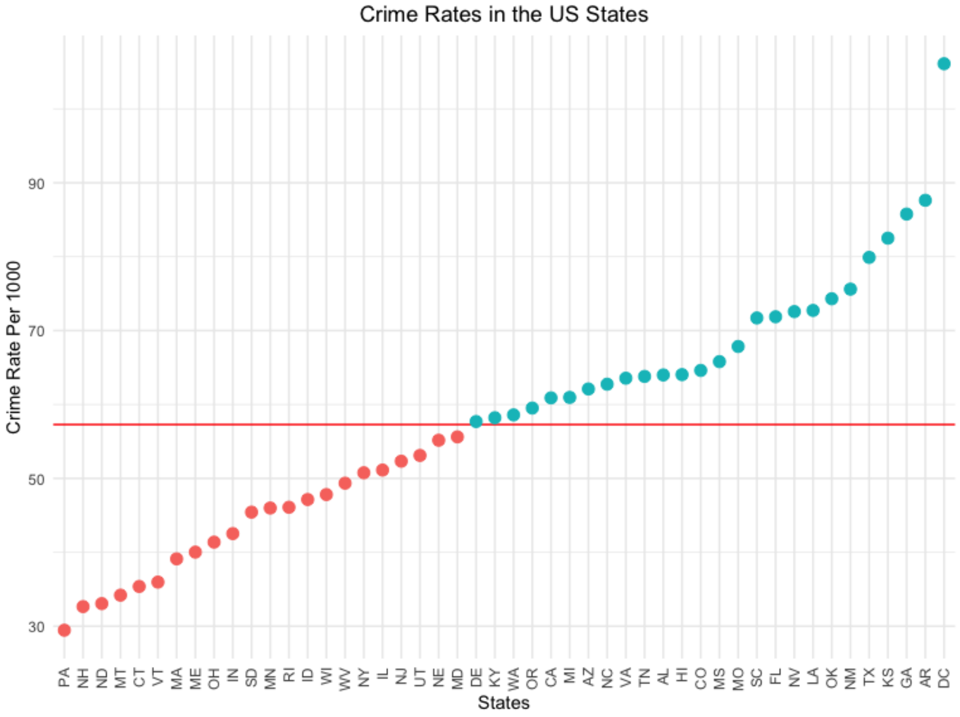


Figure 1: Overview crime rates in US

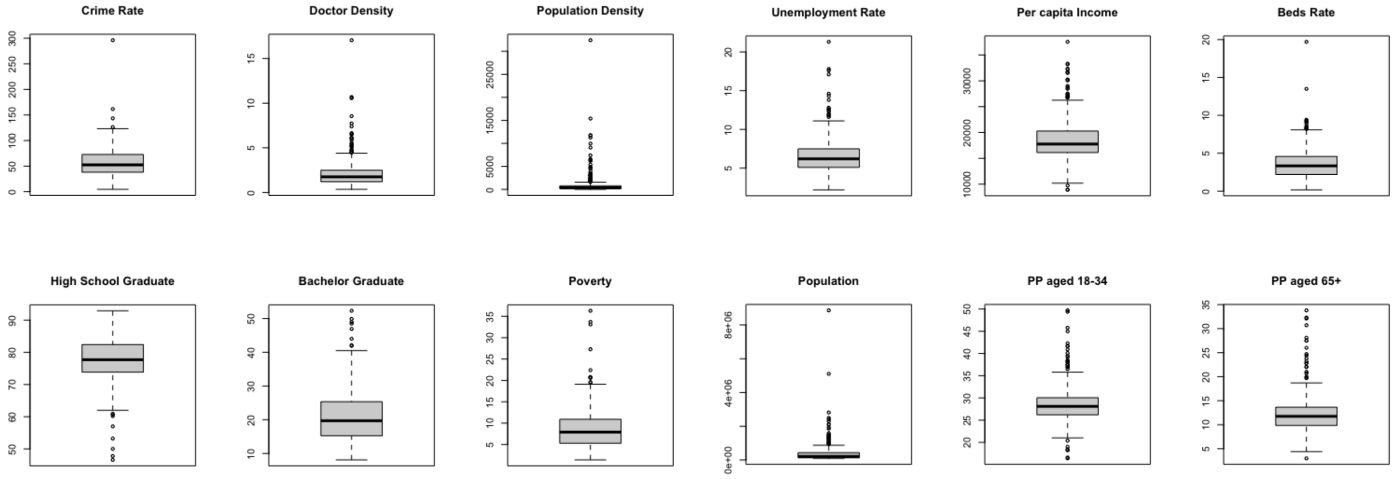
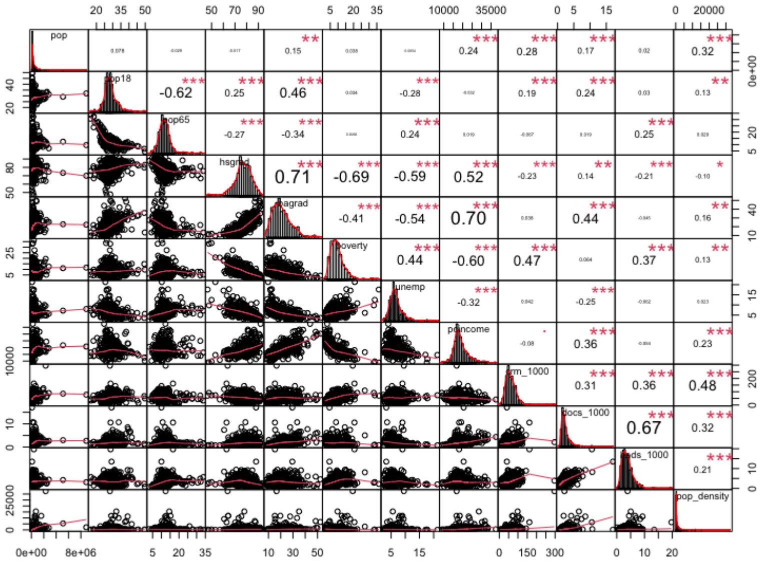


Figure 2&3: Box Plots shows the distribution of each variables



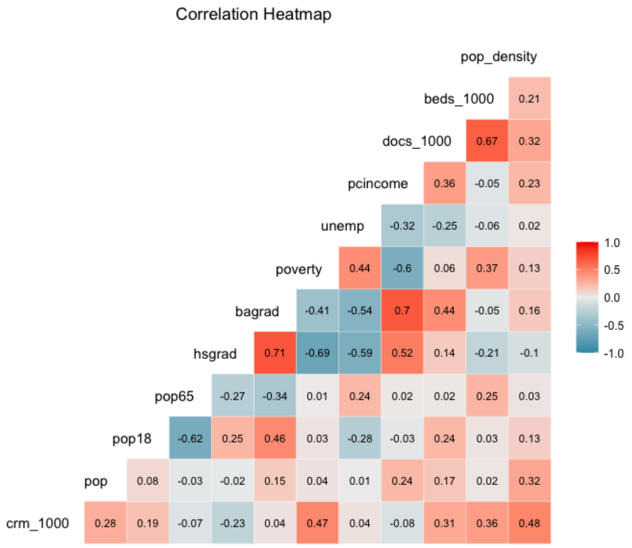


Figure 4&5: Correlation Heatmap

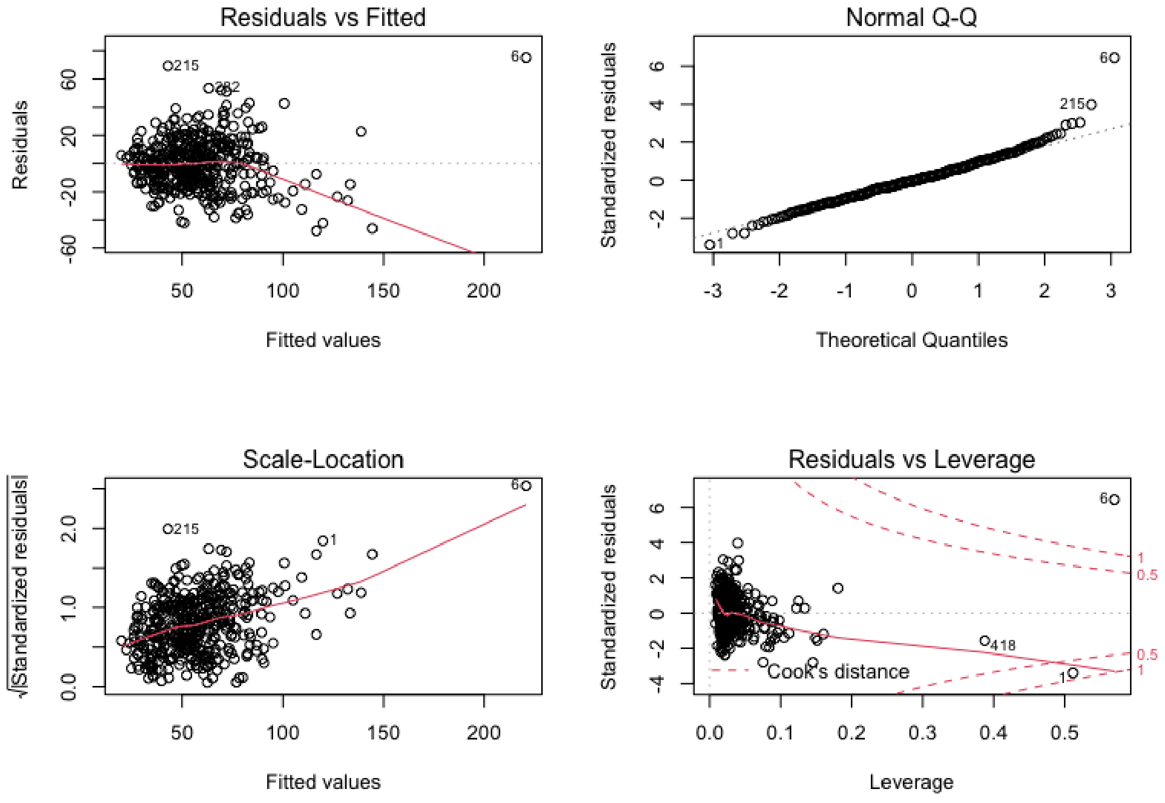


Figure 6: Full model before transformation

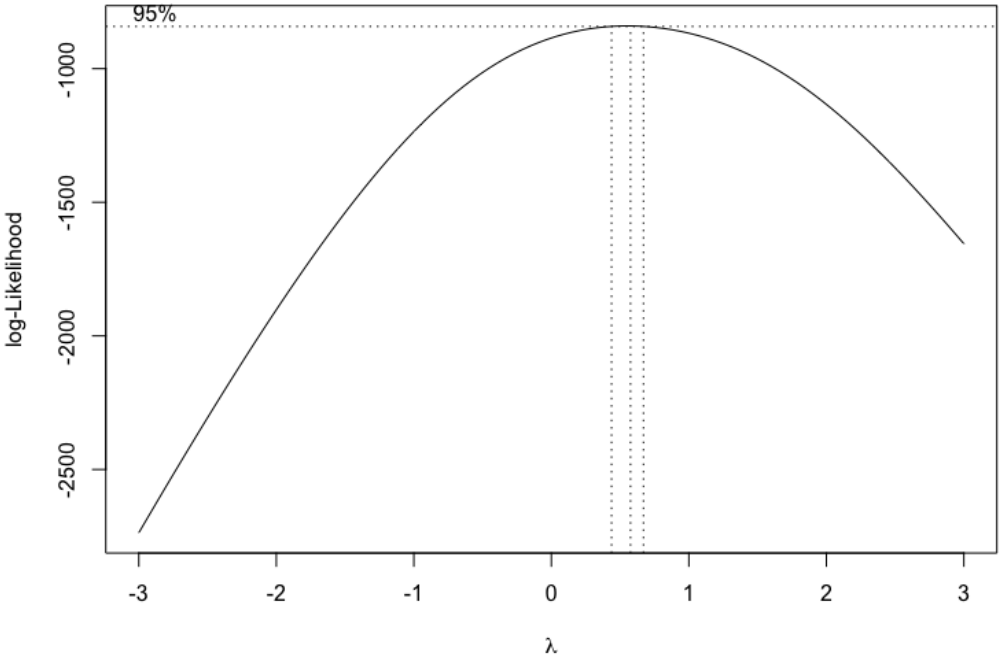


Figure 7: boxcox plot

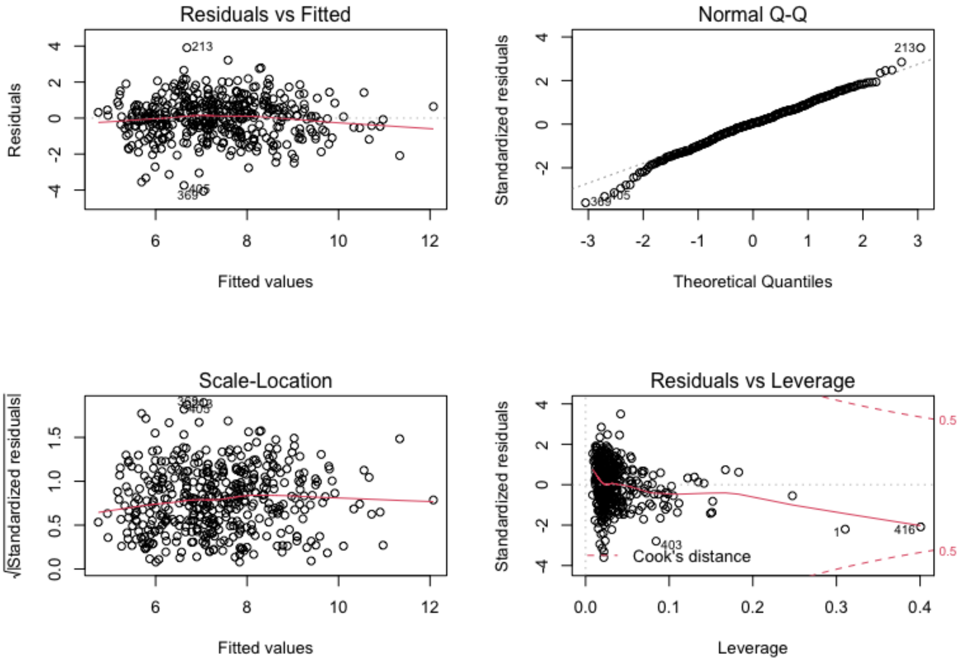
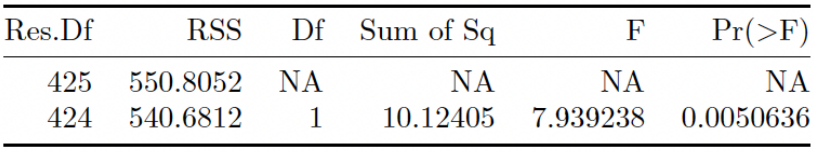


Figure 8：Full model after transformation



Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
425	550.8052	NA	NA	NA	NA
424	540.6812	1	10.12405	7.939238	0.0050636

Table 2: Partial F-test Anova Table

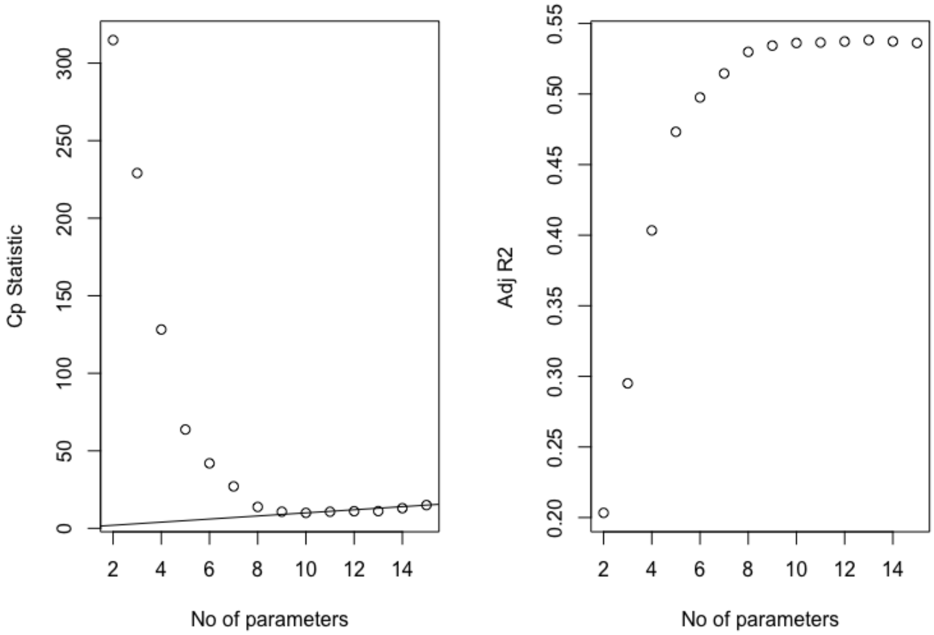
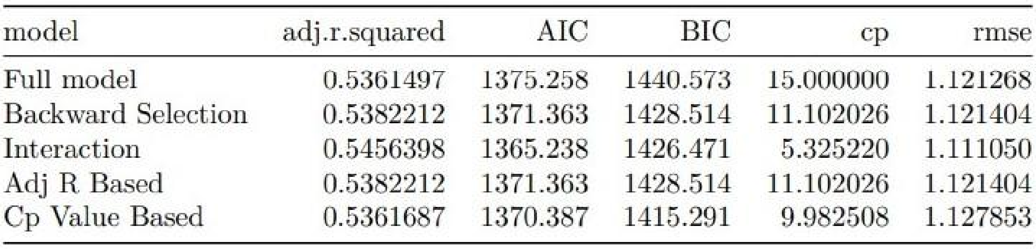


Figure 9: Cp & Adj-R^2 as functions of parameter



model	adj.r.squared	AIC	BIC	cp	rmse
Full model	0.5361497	1375.258	1440.573	15.000000	1.121268
Backward Selection	0.5382212	1371.363	1428.514	11.102026	1.121404
Interaction	0.5456398	1365.238	1426.471	5.325220	1.111050
Adj R Based	0.5382212	1371.363	1428.514	11.102026	1.121404
Cp Value Based	0.5361687	1370.387	1415.291	9.982508	1.127853

Table 3: Summary Table

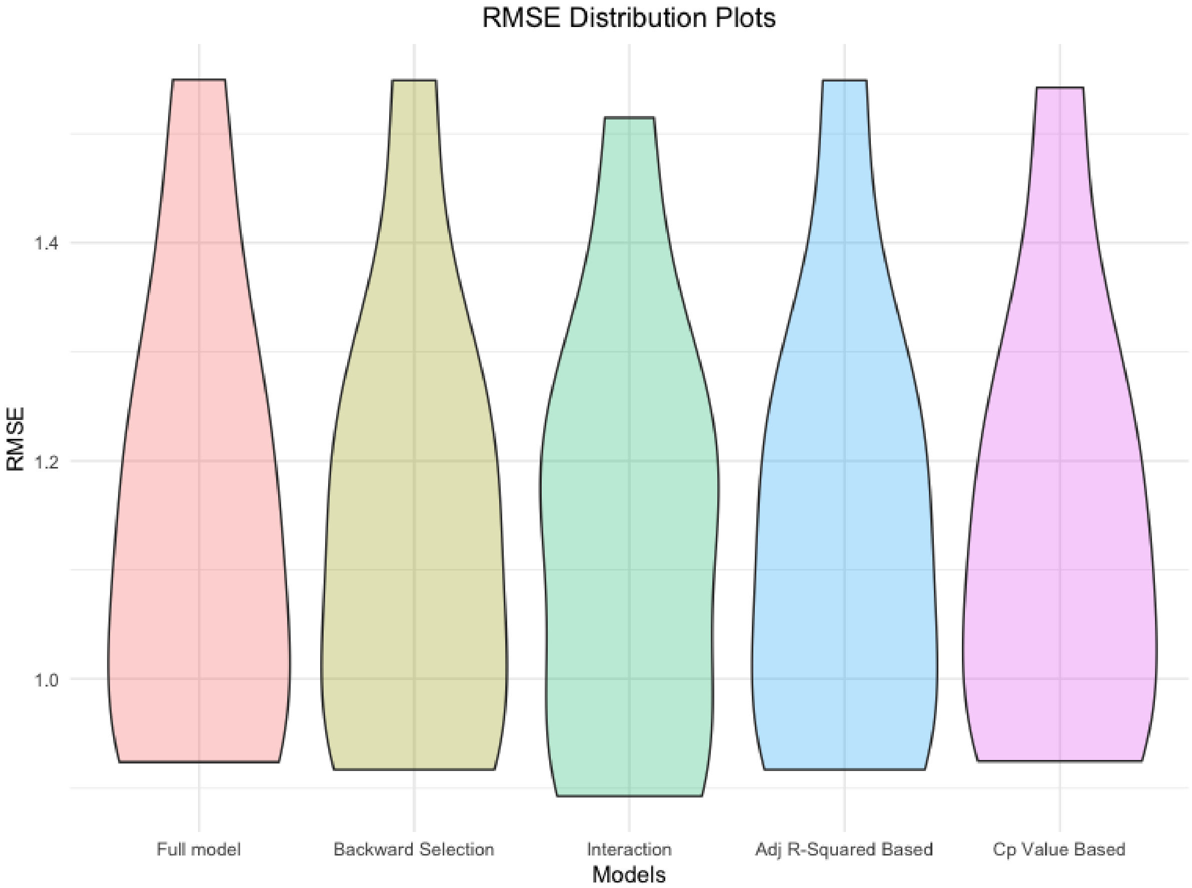


Figure 10: Cross Validation

Reference

[1] FBI. (2012, November 5). *Variables affecting crime*. FBI. Retrieved December 17, 2021, from https://ucr.fbi.gov/hate-crime/2011/resources/variables-affecting-crime