# P8106 Midterm Project Report

Ruiyang Li

3/29/2021

## Introduction

### Motivation

Public safety has always been one of the most concerns in our daily lives. A safe environment with a low crime rate not only protects us from injury and illness but also helps us improve our overall quality of life. Further, reducing crimes can lead to a decrease in societal costs and, as a result, generate substantial economic benefits. Therefore, in this project, we would like to identify the influential factors that could potentially predict the crime rate and hopefully provide insight for future policy to improve population safety. Specifically, we focus on the per capita crime rate within the Boston area. The overall goal of this project is to predict the crime rate and identify the important predicting variables as well as understand their impact on population safety.

### Data

The dataset that we used in this project can be assessed from the StatLib library which is maintained at Carnegie Mellon University. It contains information about various factors that could potentially explain crime rate from 506 observations. This dataset has already been cleaned. The detailed description of each variables is displayed in Table 1. The outcome of interest is the per capita crime rate.

We randomly split our dataset into a training dataset (75%) and a test dataset (25%). We performed model fitting and model selection on the training dataset and evaluated its test performance on the test dataset.

There were no missing or duplicated observations in our data and the distribution of each predictor variable can be seen in Figure 1. After careful consideration, we decided not to perform transformations in order to maintain the interpretability of the predictors.

## Exploratory analysis/visualization

From the scatter plot of each predictor versus crime rate (Figure 2), we can see that some variables seem to have nonlinear patterns with crime rate whereas others seem to have linear patterns. Therefore, it is reasonable to consider using non-linear models, such as GAM and MARS, to fit the data in addition to linear models.

We also examined the multi-collinearity among predictors because if two predictors are highly correlated, it may undermine the statistical significance of the variable. From the correlation plot (Figure 3), we can observe that there are a number of predictors highly correlated with each other. For example, the per capita crime rate (CRIM) seems to have a high positive correlation with the index of accessibility to radial highways (RAD) and full-value property-tax rate per $10,000 (TAX). The weighted distances to five Boston employment centers (DIS) seem to have a high positive correlation with the proportion of residential land

zoned for lots over 25,000 sq.ft. (ZN) and a high negative correlation with the proportion of non-retail business acres per town (INDUS), nitrogen oxideS concentration per 10 million (NOX), and proportion of owner-occupied units built prior to 1940 (AGE). The proportion of non-retail business acres per town (INDUS) and % lower status of the population (LSTAT) seem to be correlated with most predictors except for CHAS.

We will explore more about the predicting abilities of these variables for crime rate in the Models section.

## Models

We used crime rate as our response variable and all the remaining variables as the predictors. Since crime rate is a continuous outcome, we performed the modeling under the regression framework. Our earlier exploratory findings from the scatter plot suggested that there seemed to be some non-linear association between the predictors and the outcome, therefore, we considered both linear and non-linear models during the modeling process. Specifically, we fitted linear models including multiple linear regression, LASSO, ridge, elastic net, principal component regression (PCR), and partial least squares (PLS) model, and non-linear models such as generalized additive model (GAM) and multivariate adaptive regression splines (MARS).

There were one or few tuning parameters for each of the models except the linear regression model. Within these models that require tuning parameters, we specified a grid of numbers and then selected the one that gave the smallest cross-validation error. We then compared each of these models with optimal tuning parameters and decided the final model based on their training performance, i.e., training RMSE. Both median and mean RMSEs are reasonable evaluation metrics for the selection of the final model. We discussed both metrics and selected the model with the smallest training RMSE. And finally, the final selected model would be used to predict the crime rate.

Figure 4 displays the distribution of the training errors (in black) of each of the models that we considered in this project. We can see that the MARS model has both the smallest mean and median training RMSEs. Hence, we decided to use it as our final model.

We also evaluated the test performance of the MARS model in the form of mean RMSE as displayed in Figure 4 (in orange). We can observe that MARS has the smallest test error among all the models and its test error is close to its training error. This is desirable and indicates that MARS seems to have a consistently reliable predictive ability. Therefore, our decision of choosing MARS looks reasonable and MARS seems to be flexible enough and able to successfully capture the non-linearity between the predictors and the crime rate. By the nature of MARS, no specific assumption is made in MARS with regard to the functional relationship between the response variable and the predictor variables.

Our final model, the MARS model, has the following form.

$$\begin{aligned}\hat{CRIM} = {} & 0.122 + 0.524\ h(RAD - 8) \\ & + 0.786\ h(2.20 - DIS) * h(RAD - 8) \\ & + 0.309\ h(23.24 - LSTAT) * h(17.4 - MEDV) \\ & - 0.020\ h(RAD - 8) * h(MEDV - 13.8) \\ & + 0.125 h(RAD - 8) * h(13.8 - MEDV) \\ & - 0.135\ h(85.5 - AGE) * h(17.4 - MEDV) \\ & + 30.5\ h(NOX - 0.7) * h(RAD - 8) \\ & - 25.4\ h(NOX - 0.679) * h(RAD - 8)\end{aligned}$$

Based on the variable importance plot (VIP) in Figure 5 (first plot), RAD, DIS, MEDV, LSTAT, and AGE are the top 5 variables of importance in predicting the crime rate. The partial dependence plots (PDP) (Figure 5) illustrate how these variables impact the prediction of the response variable. For example, PDP of RAD implies that our model found that one knot (at 8) in RAD provides the best fit. Specifically, the per

capita crime rate seems to be stable when the index of accessibility to radial highways is below 8 but it starts to increase after the index exceeds 8. Similarly, we can see that the crime rate seems to first decrease with weighted distances to five Boston employment centers (DIS) and remain stable after around 2.20. Overall, the crime rate seems to decrease for the median value of owner-occupied homes, but it decreases fastest when the median value is below 13.8, second fastest when the median value is from 13.8k to 17.4k, and the slowest when the median value exceeds 17.4k. The crime rate seems to decrease with the percentage of the lower status of the population and remaining stable after around 23.24%.

# Discussion

We did not perform transformations on the predictors due to the concern of interpretability. Through transformation might compromise the interpretability, it could possibly improve model performance. Future research could consider applying transformation on predictors.

Overall, our model suggests that accessibility to radial highways, distances to employment centers, and the median value of owner-occupied homes are the top 3 variables that play an important role in predicting the crime rate. Specifically, the crime rate is predicted to start to increase with the accessibility to radial highways when the index of latter exceeds 8, to decrease if the distance to employment centers is farther, and to decreases with the median home value overall – it decreases faster if the median is below 17.4k compared to after 17.4k. These findings are consistent with our natural understanding and expectation and provide meaningful insight into public safety. Therefore, the policymaker or authority might consider making more effort in these forehead mentioned areas to reduce the crime rate and protect the overall population.

## Tables and Figures

| Variable | Description |
|---|---|
| AGE | proportion of owner-occupied units built prior to 1940 |
| B | 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town |
| CHAS | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) |
| CRIM | per capita crime rate by town |
| DIS | weighted distances to five Boston employment centres |
| INDUS | proportion of non-retail business acres per town |
| LSTAT | % lower status of the population |
| MEDV | Median value of owner-occupied homes in $1000s |
| NOX | nitrogen oxides concentration (parts per 10 million) |
| PTRATIO | pupil-teacher ratio by town |
| RAD | index of accessibility to radial highways |
| RM | average number of rooms per dwelling |
| TAX | full-value property-tax rate per $10,000 |
| ZN | proportion of residential land zoned for lots over 25,000 sq.ft. |

**Table 1** Variable Description

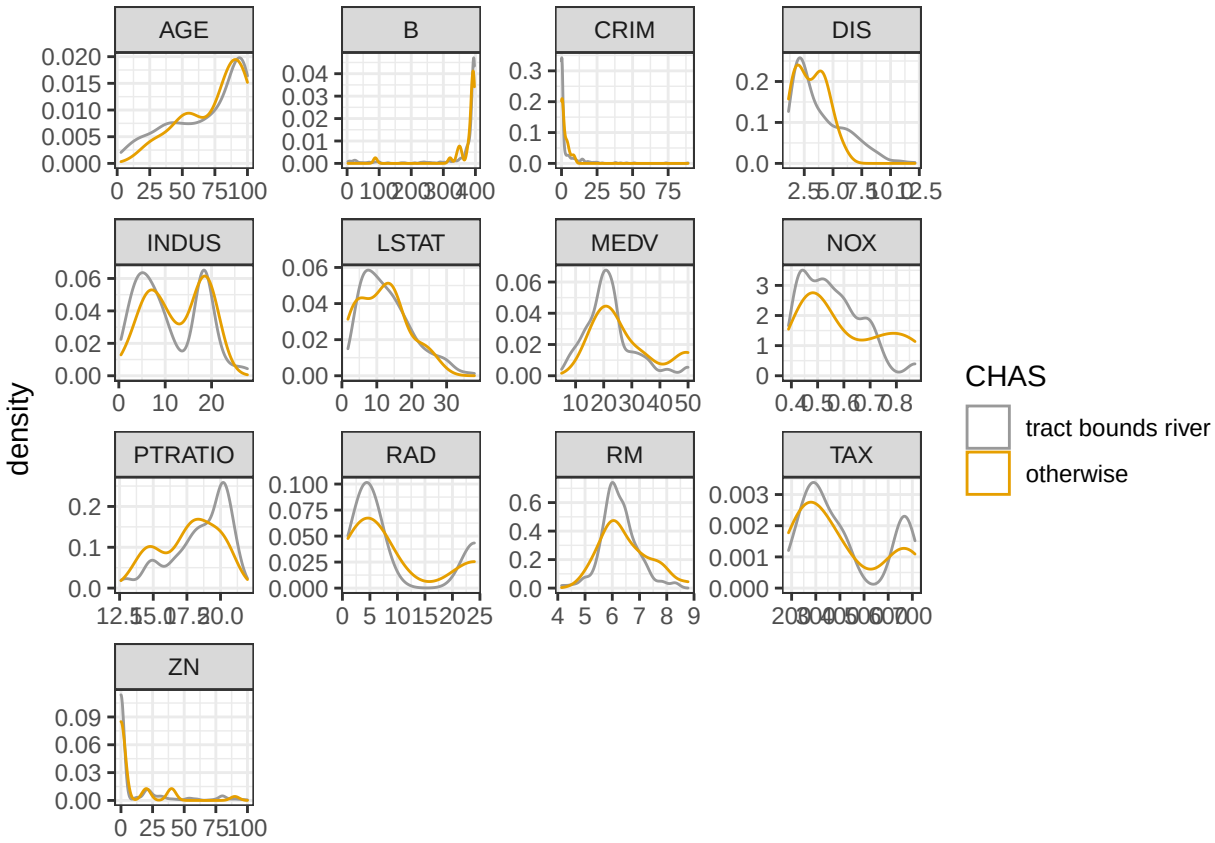


**Figure 1** Variable Distribution (training dataset)

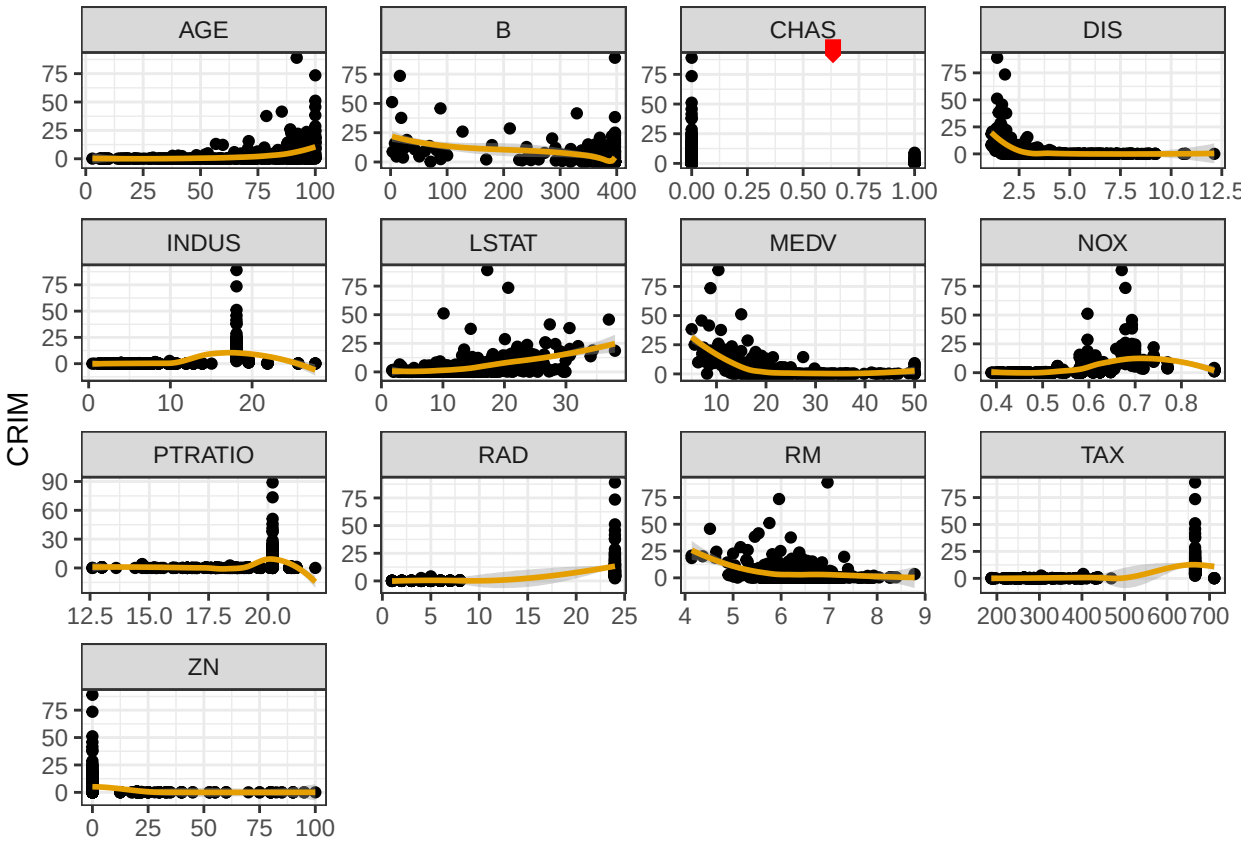


**Figure 2** Scatter plot of each variable vs. crime rate (training dataset)

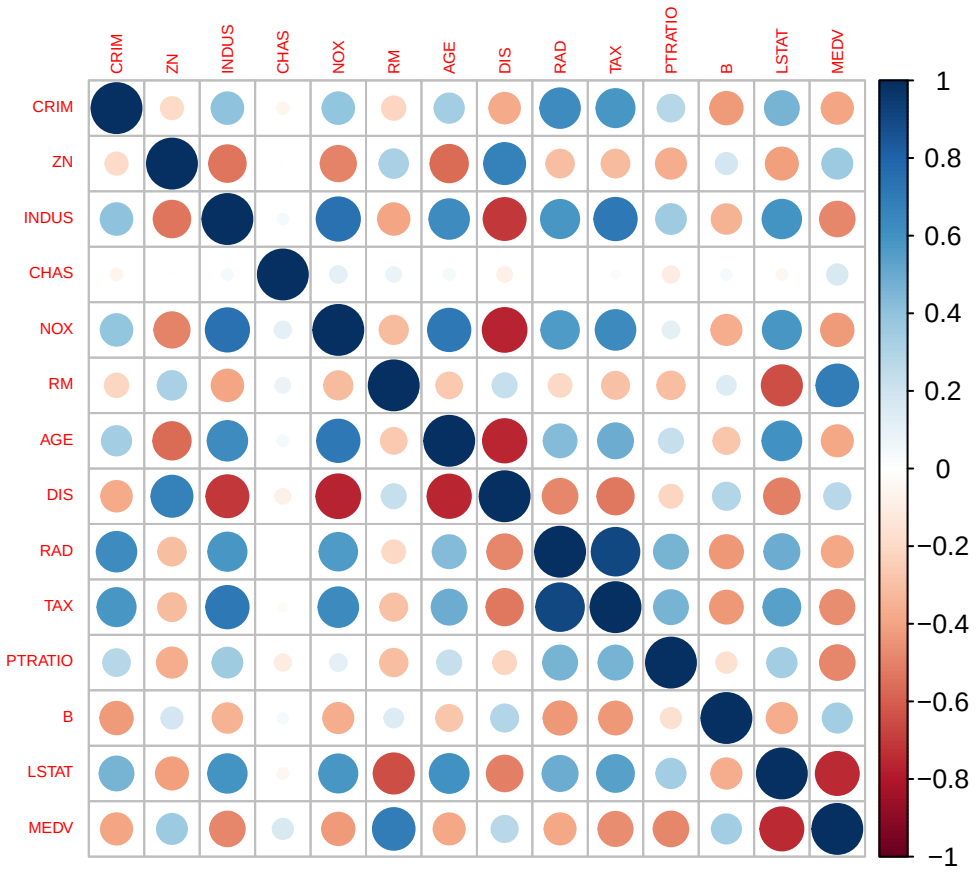


**Figure 3** Correlation Plot (training dataset)

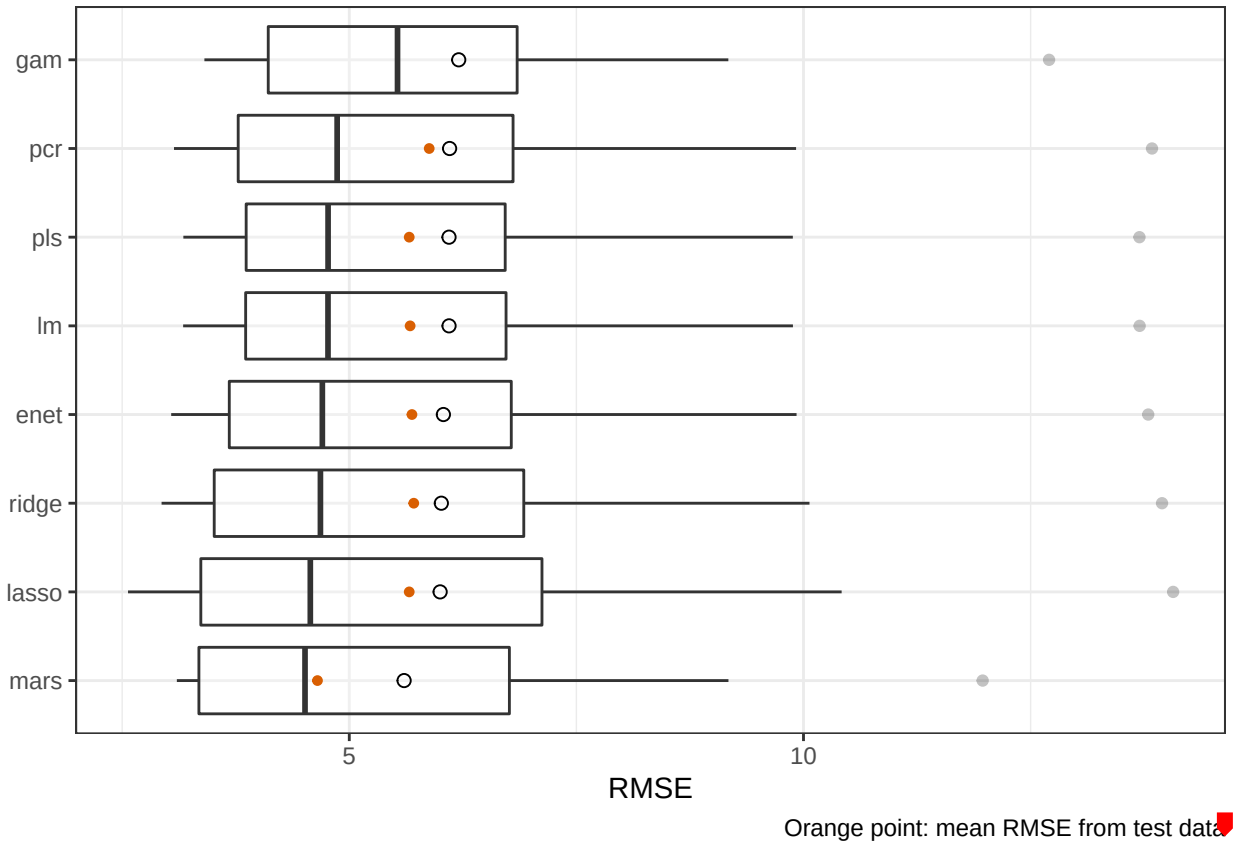


**Figure 4** Predictive performance of all models

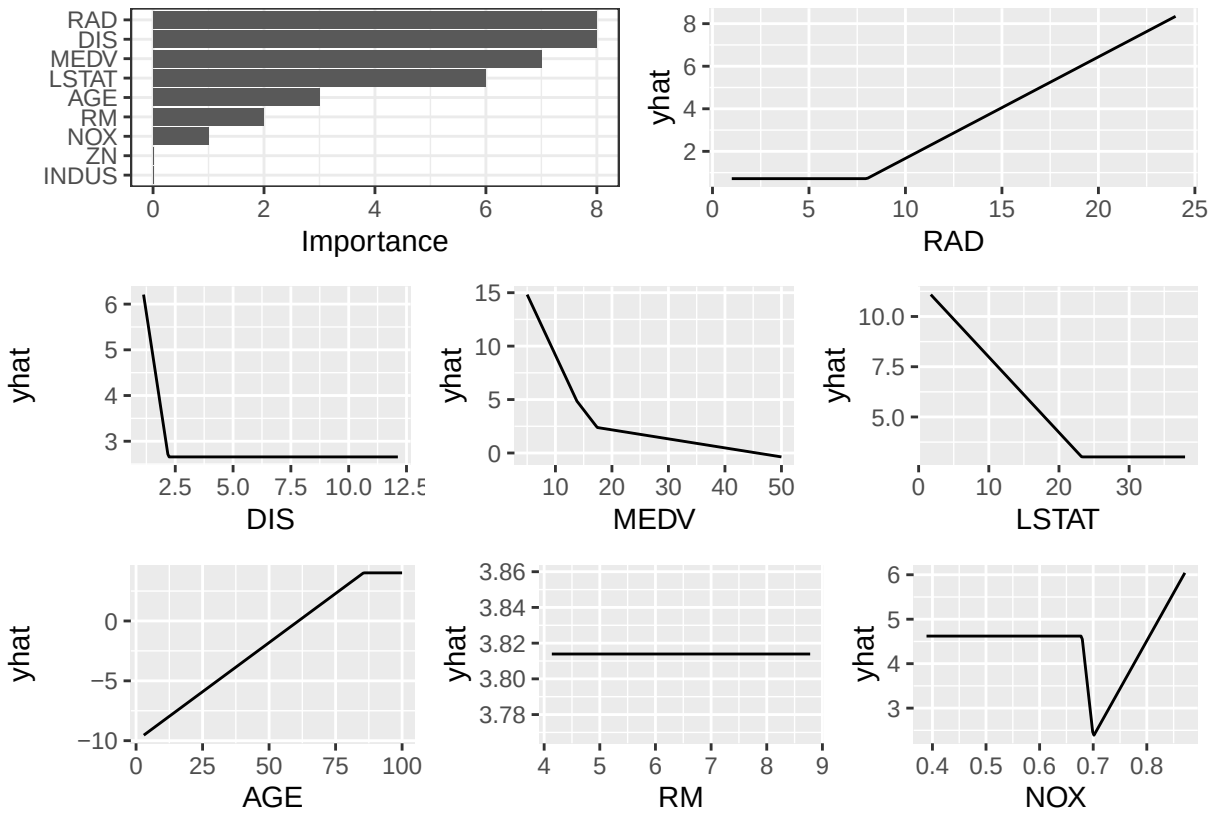


**Figure 5** The Variable Importance Plot (VIP) and Partial Dependence Plots (PDP) from the MARS model (final model)