Predicting Crime Rate Within Boston Area

Group 20

5/13/2021

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 14 variables that could potentially explain crime rate from 506 observations. 13 out of 14 variables are continuous variables, and the variable CHAS is a dichotomous variable, which has been converted to the factor type. The detailed description of each variables is displayed in Table 1. The outcome of interest is the per capita crime rate (CRIM), and the remaining 13 variables are the predictors.

This dataset has already been cleaned and does not contain missing or duplicated values. If the variable names are too messy, we can use the `janitor::clean_names()` function to make the variable names unique and consist only of the "_" character, numbers, and letters. If there are missing values, we may consider using r functions like `na.omit` or `drop_na()` to remove them. 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.

We used `createDataPartition()` function to 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.

Exploratory Analysis/Visualization

From the 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 (e.g., GAM and MARS), tree-based methods (e.g., regression tree and conditional inference tree), and ensemble methods (e.g., random forests and boosting) to fit the data in addition to linear models.

We also examined the multi-collinearity among predictors. This is 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 following Models section.

Models

Model Determination

There are no missing values among all the variables. We checked through all the variables to see if there is any near zero variance predictor using the `nearZeroVar` function in the datasets before and after splitting, and the result shows there does not exist near zero variance predictors. Therefore, 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 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), multivariate adaptive regression splines (MARS), regression tree, conditional inference tree, random forests, and boosting.

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., cross-validation RMSE. Both median and mean RMSEs are reasonable evaluation metrics for the selection of the final model. We discussed both metrics and determined the appropriate final model, which would be used to predict the crime rate.

Model Tuning

(1) Linear Regression: Since Linear Model has a closed form solution, we did not need to tune parameters for this model. At 5% significance level, there are 5 out of 13 predictors that are significant, they are ZN, DIS, RAD, B, and MEDV.

(2) Ridge Regression: We tuned 100 lambda values in between e^2 to e^{-3}. From the Figure A1 in the appendix, the lambda value with the smallest cross-validation RMSE is 2.2, with a corresponding log value 0.79 on the plot. The Ridge model with the smallest cross-validation RMSE utilized all 13 predictors.

(3) Lasso Regression: We tuned 100 lambda values in between e to e^{-5}. From the Figure A2 in the appendix, the best lambda value with the smallest cross-validation RMSE is 0.81, with a corresponding log value -0.21 on the plot. The Lasso model with the smallest cross-validation RMSE utilized 4 predictors, which are RAD, B, LSTAT, and MEDV.

(4) Elastic Net Model: We tuned 5 alpha values in between 0 to 1, and tuned 50 lambda values in between e^{-10} to e^{-3}. From the Figure A3 in the appendix we can see alpha = 0 (ridge model) has significant lower cross-validation RMSE compared with other alpha values. The Elastic Net Model with the smallest cross-validation RMSE utilized all 13 predictors.

(5) Principle component regression (PCR) model: We tuned the number of components from 1 to 13. The Figure A4 in the appendix shows the optimal components number is 13.

(6) Partial Least Squares (PLS) Model: We tuned the number of components from 1 to 13 and the Figure A5 in the appendix shows the optimal components number is 10.

(7) Generalized Additive Model (GAM): The GAM model did not provide us the way to tune the parameters. The function automatically returned the most optimal model. The best GAM model is: outcome ~ CHASotherwise + RAD + s(ZN) + s(PTRATIO) + s(TAX) + s(INDUS) + s(NOX) + s(MEDV) + s(B) + s(AGE) + s(DIS) + s(RM) + s(LSTAT).

(8) Multivariate Adaptive Regression Splines (MARS): We tuned the maximum degree of interaction (degree) from 1 to 3 and the maximum number of terms (including intercept) in the pruned model (nprune) from 2 to 22. The best model with the smallest cross-validation RMSE is with nprune = 9 and degree = 2. (see Figure A6 in the appendix) The best model Selected 9 of 22 terms, and 6 of 13 predictors (DIS, RAD, MEDV, LSTAT, AGE, and NOX).

(9) Regression Tree Model: We tuned 50 values of Complexity Parameters (cp, the minimum improvement in the model needed at each node) from e^{-6} to e^{-3}. The model with the smallest cross-validation RMSE is with cp = 0.01376379. (see Figure A7 in the appendix)

(10) Conditional Inference Tree Model: We tuned 50 values of the minimum criterion for a split (mincriterion) from $1-e^{-5}$ to $1-e^{-3}$. The model with the smallest cross-validation RMSE is with mincriterion = 0.9779915. (see Figure A8 in the appendix)

(11) Random Forest Model: We tuned the number of selected variable at each split from 1 to 13, and tuned the minimum node size from 1 to 6. The model with the smallest cross-validation RMSE is with number of selected variable at each split equal to 1 and minimum node size equal to 4. (see Figure A9 in the appendix)

(12) Boosting Model: We tuned the number of trees (n.trees) with values equal to 1000, 2000 and 3000, tuned the maximum depth of each tree (interaction.depth) with values from 1 to 10, tuned the shrinkage parameter with values equal to 0.0008, 0.001, and 0.0012, and tuned the minimum number of observations in terminal nodes (n.minobsinnode) from 13 to 16. The model with the smallest cross-validation RMSE is with n.trees = 2000, interaction.depth = 6, shrinkage = 0.001, and n.minobsinnode = 15. (see Figure A10 in the appendix)

Model Summary

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 both random forest (rf) model and boosting (gbm) model performed well compared to the other models. Both models have similar performance using the training data; the random forest model has a slightly larger mean RMSE but a smaller median RMSE than the boosting model. As there is extreme value in RMSE, median RMSE can be a more appropriate criterion as it is insensitive to extreme values whereas mean is. Hence, we decided to use the random forest model, which has a smaller median RMSE, as our final model.

We also evaluated the test performance of the models in the form of RMSE as displayed in Figure 4 (in orange). We can observe that among all the models, regression tree (rpart) model has the smallest test error, followed by boosting model (gbm), MARS model (mars), and random forest model (rf). Our selected final model, the random forest model, has its test error (RMSE) close to its training error (mean RMSE). This is desirable and it indicates that random forest model seems to have a consistently reliable predictive ability. Therefore, our decision of choosing random forest model looks reasonable and random forest seems to be flexible enough to make the prediction and able to successfully capture the non-linearity between the predictors and the crime rate.

Final Model Interpretation

By the nature of random forest model, no specific assumption is made with regard to the functional relationship between the response variable and the predictor variables. As blackbox models, the random forest model does not have an explicit form. However, it can be interpreted, both globally and locally, using visualization tools.

[**TO BE MODIFIED** 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. Though transformation might compromise the interpretability, it could possibly improve model performance. Future research could consider applying transformation on predictors.

[**TO BE MODIFIED** 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

Table 1. Variable Description

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.

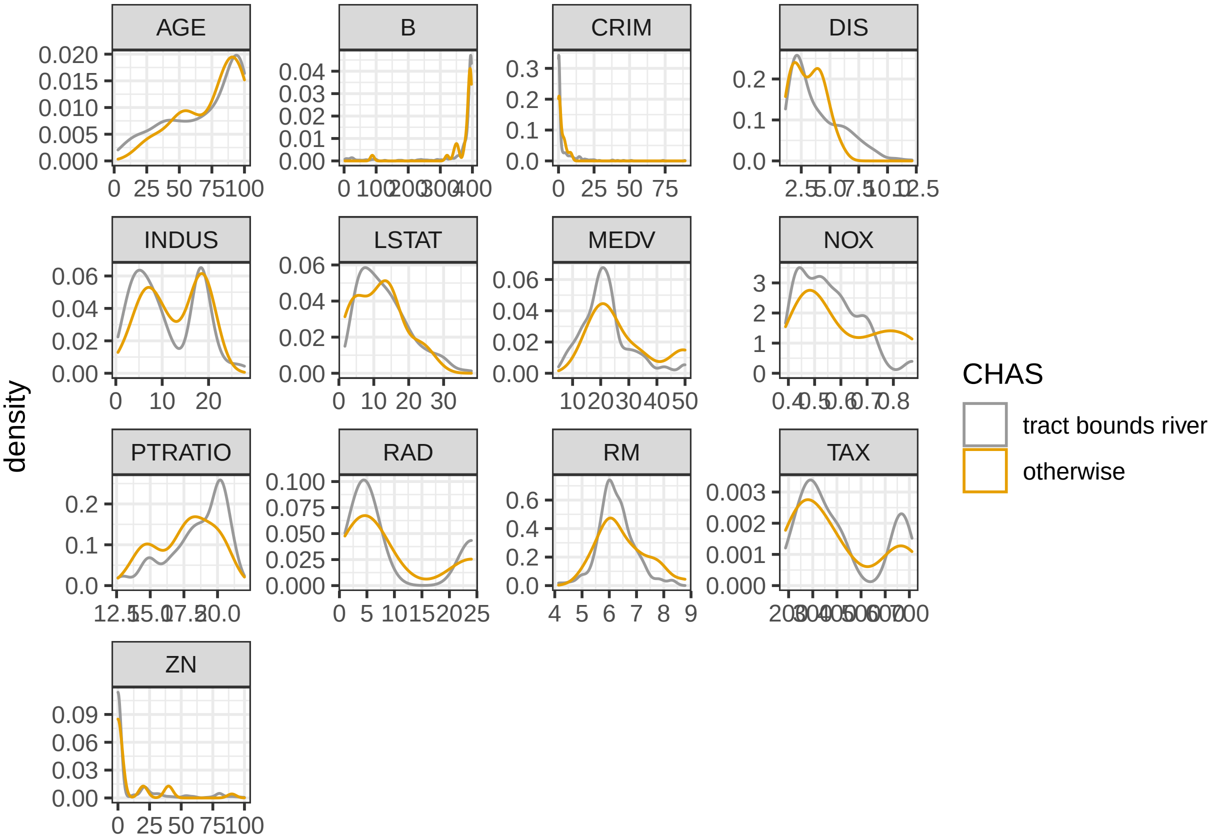

Figure 1. Variable Distribution (Training Dataset)

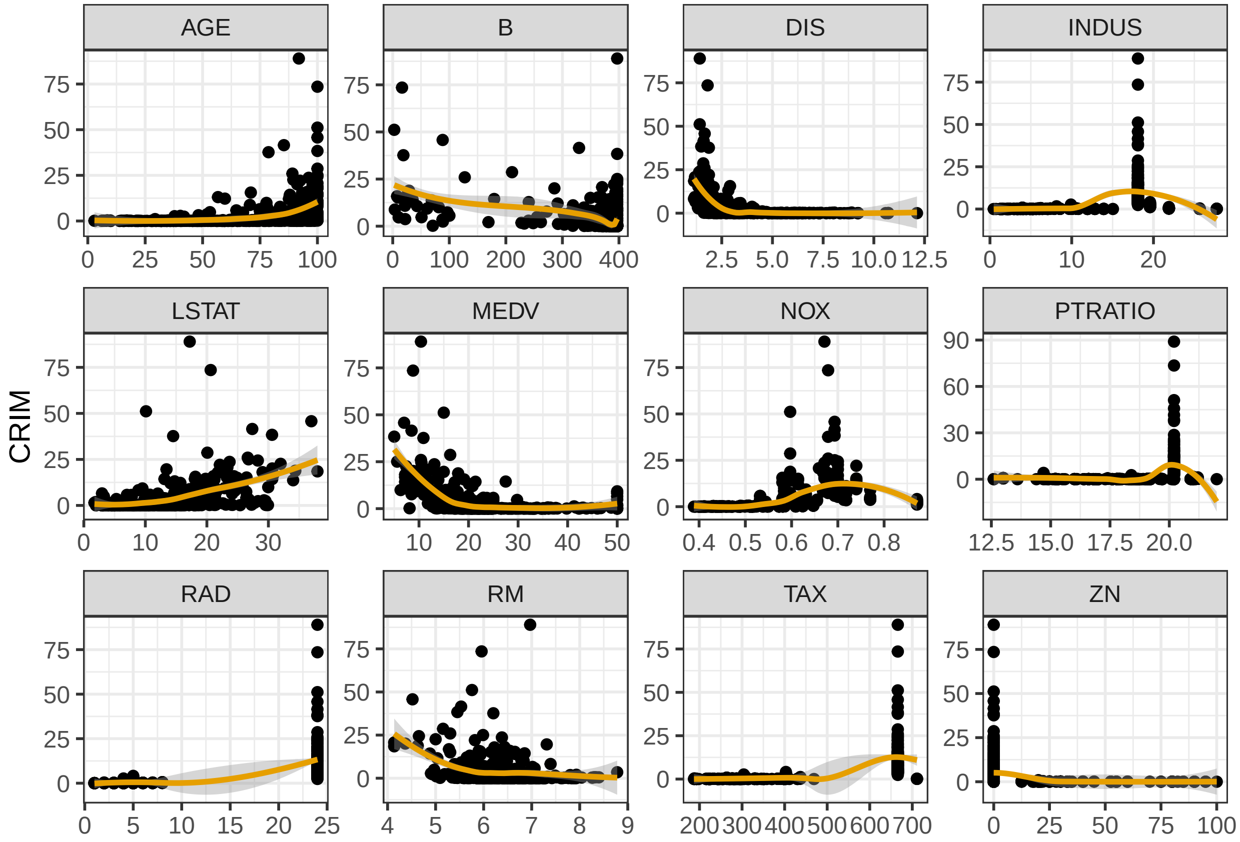

Figure 2A. Scatter Plot of Continuous Variables vs. Crime Rate (Training Dataset)

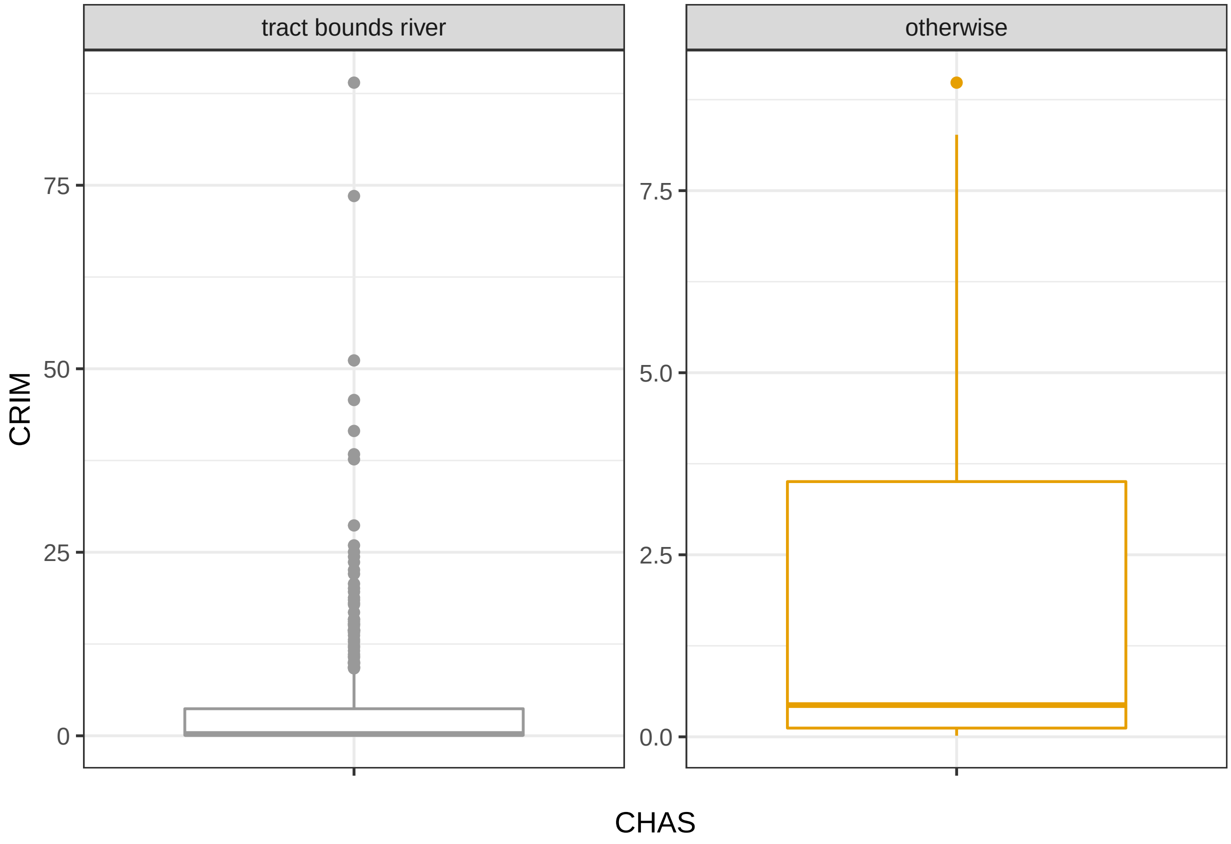

Figure 2B. Boxplot of the Categorical Variable CHAS vs. Crime Rate (Training Dataset)

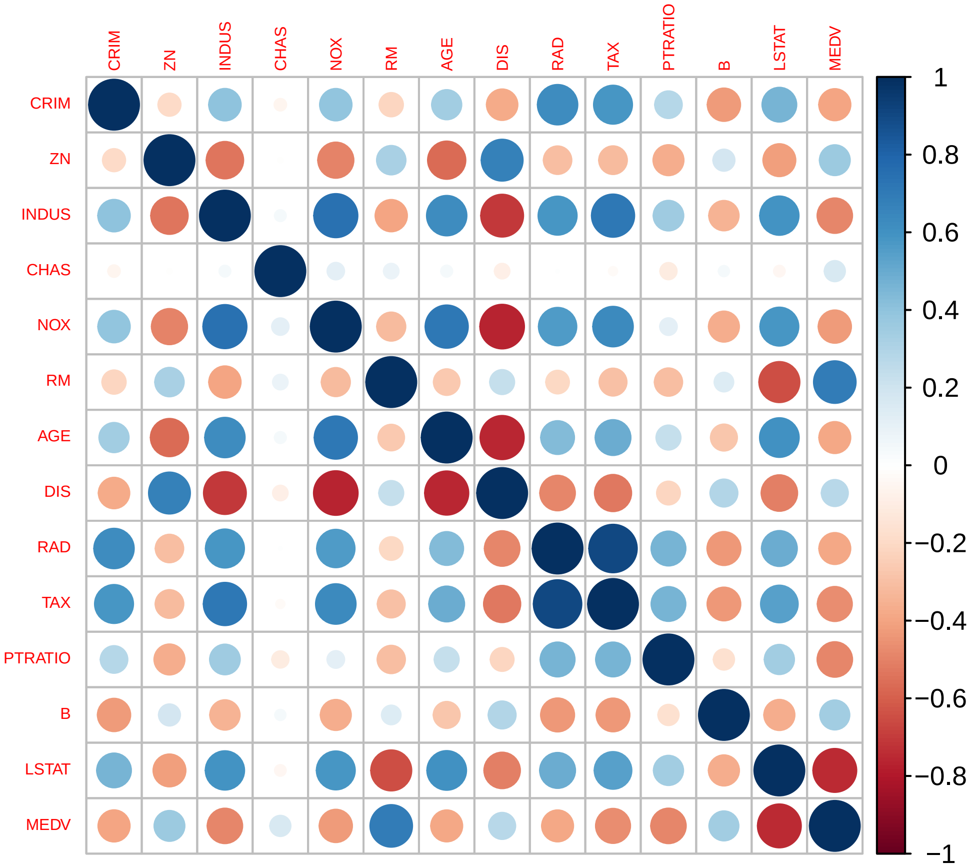

Figure 3. Correlation Plot (Training Dataset)

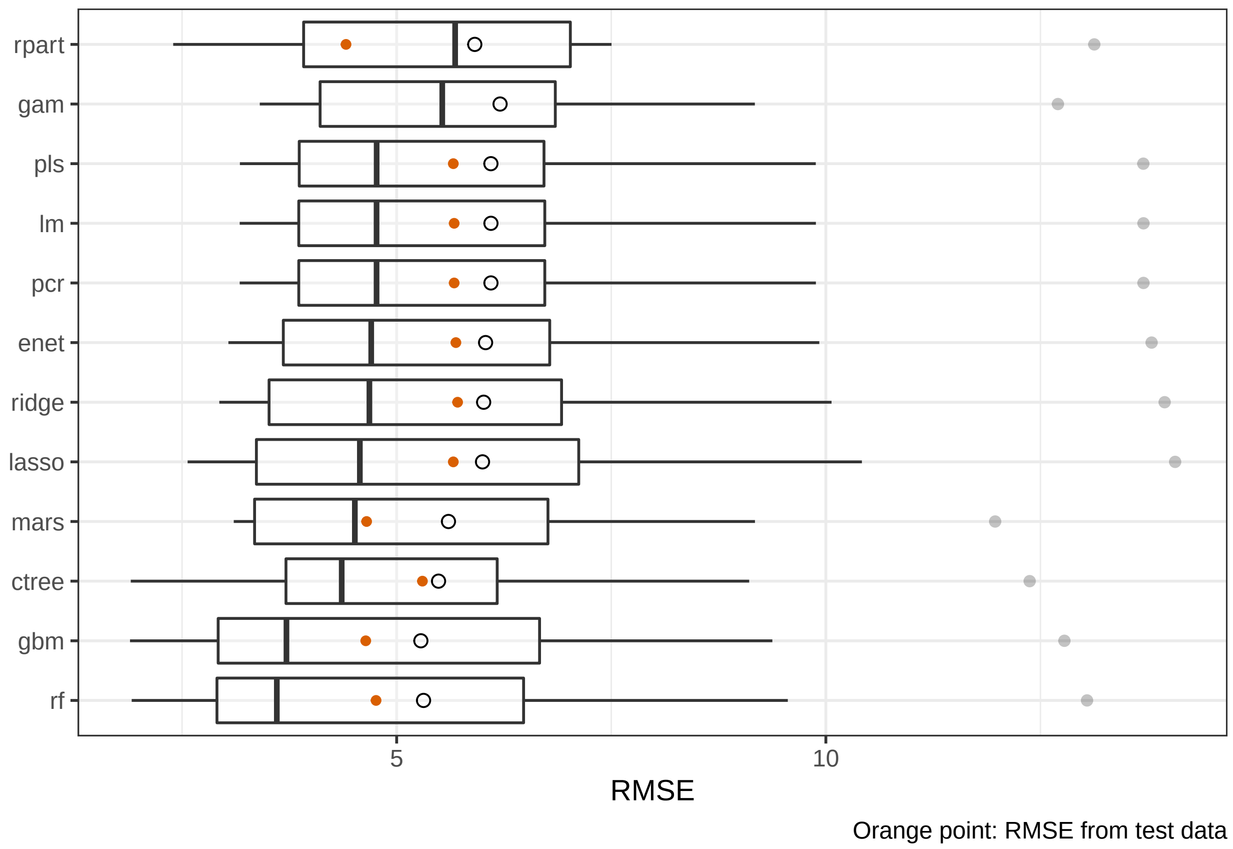

Figure 4. Predictive Performance of All Models

TO BE ADDED

Figure 5. The Variable Importance Plot (VIP), Partial Dependence Plots (PDP), Individual Conditional Expectation (ICE) Curves from Random Forest Model (Final Model)

Appendix of Model Tuning

Ridge Model

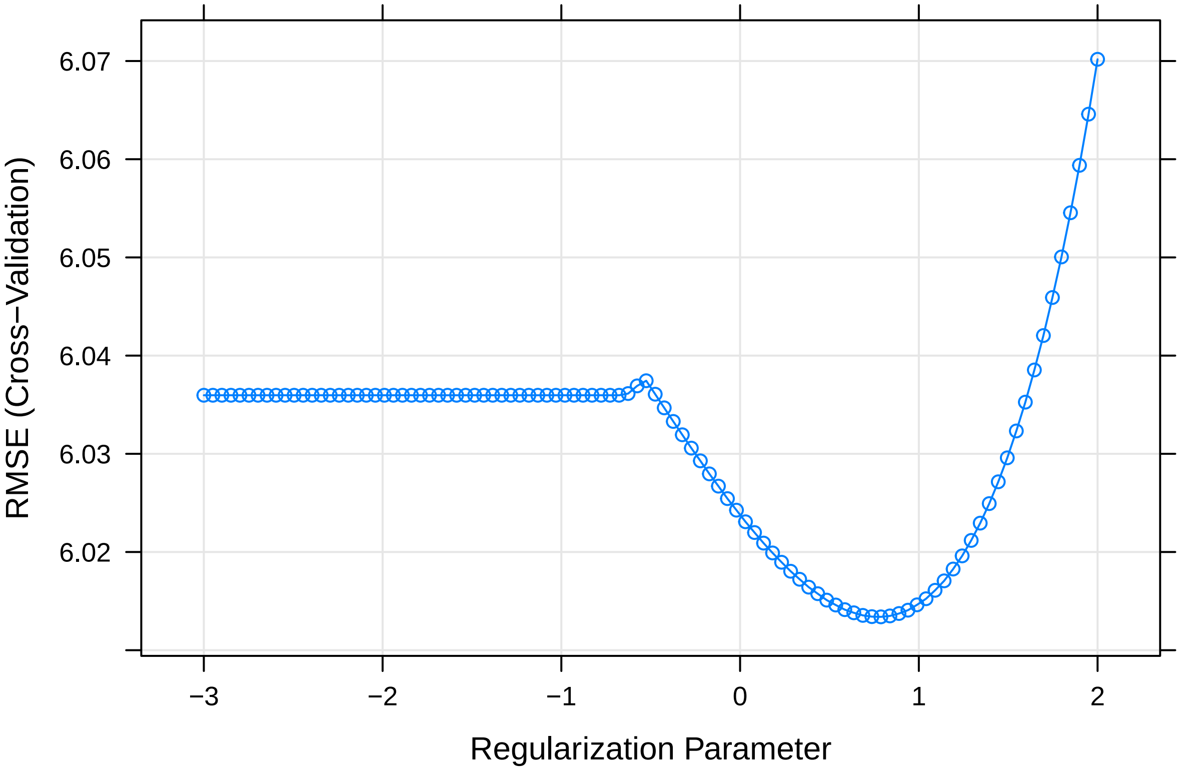

Figure A1. Tuning Lambda of Ridge Model

Lasso Model

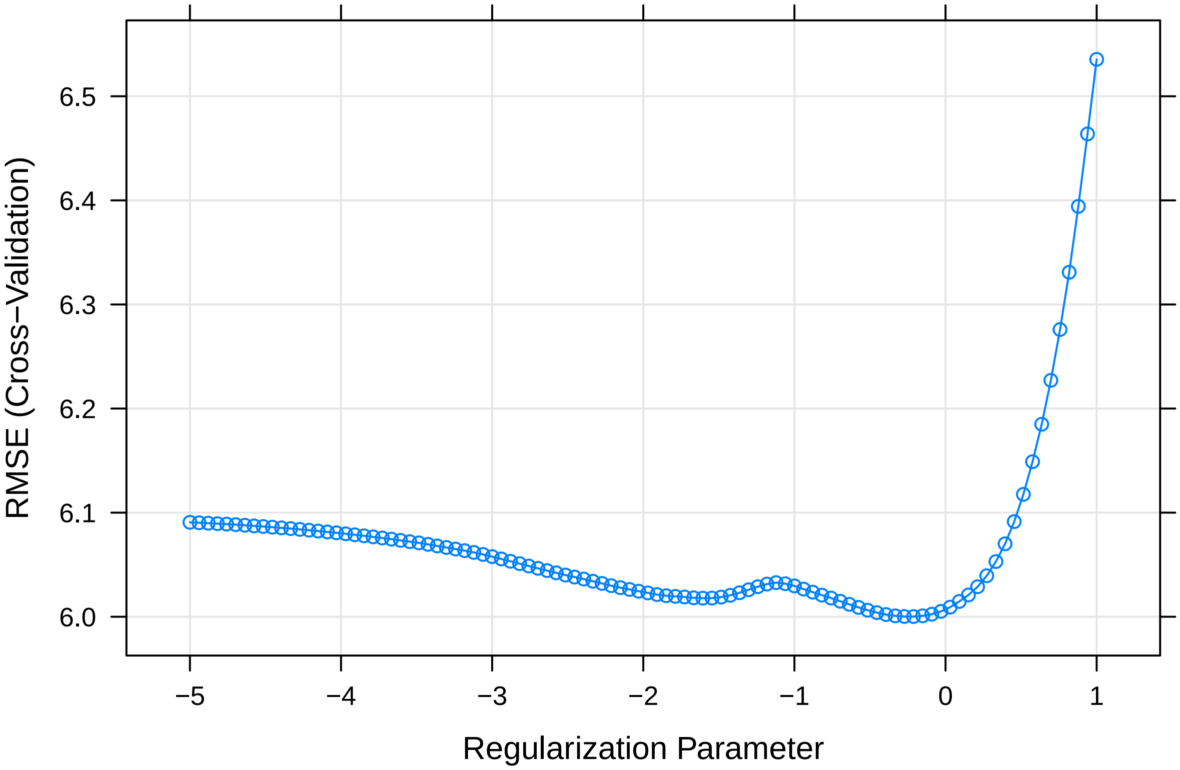

Figure A2. Tuning Lambda of Lasso Model

Elastic Net Model

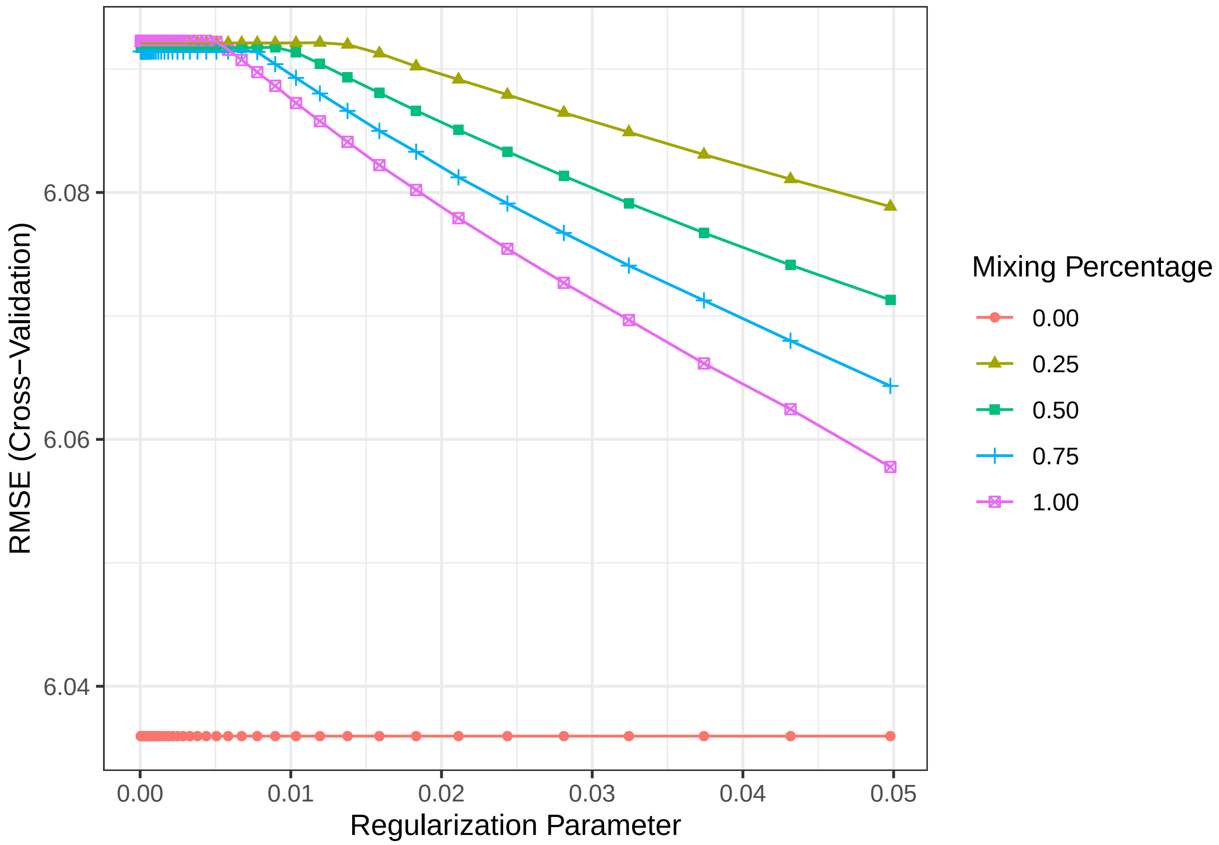

Figure A3. Tuning Alpha and Lambda of Elastic Net Model

PCR Model

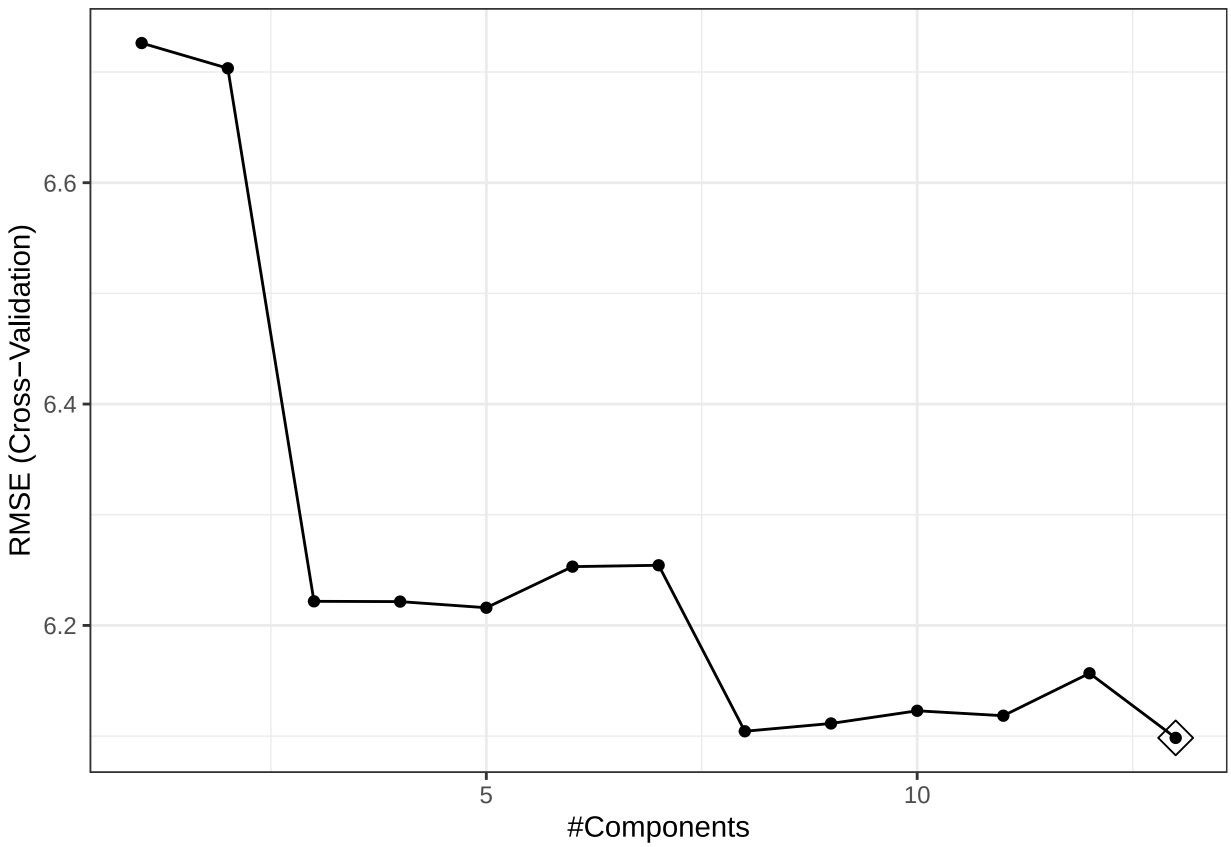

Figure A4. Tuning Optimal Components of PCR Model

PLS Model

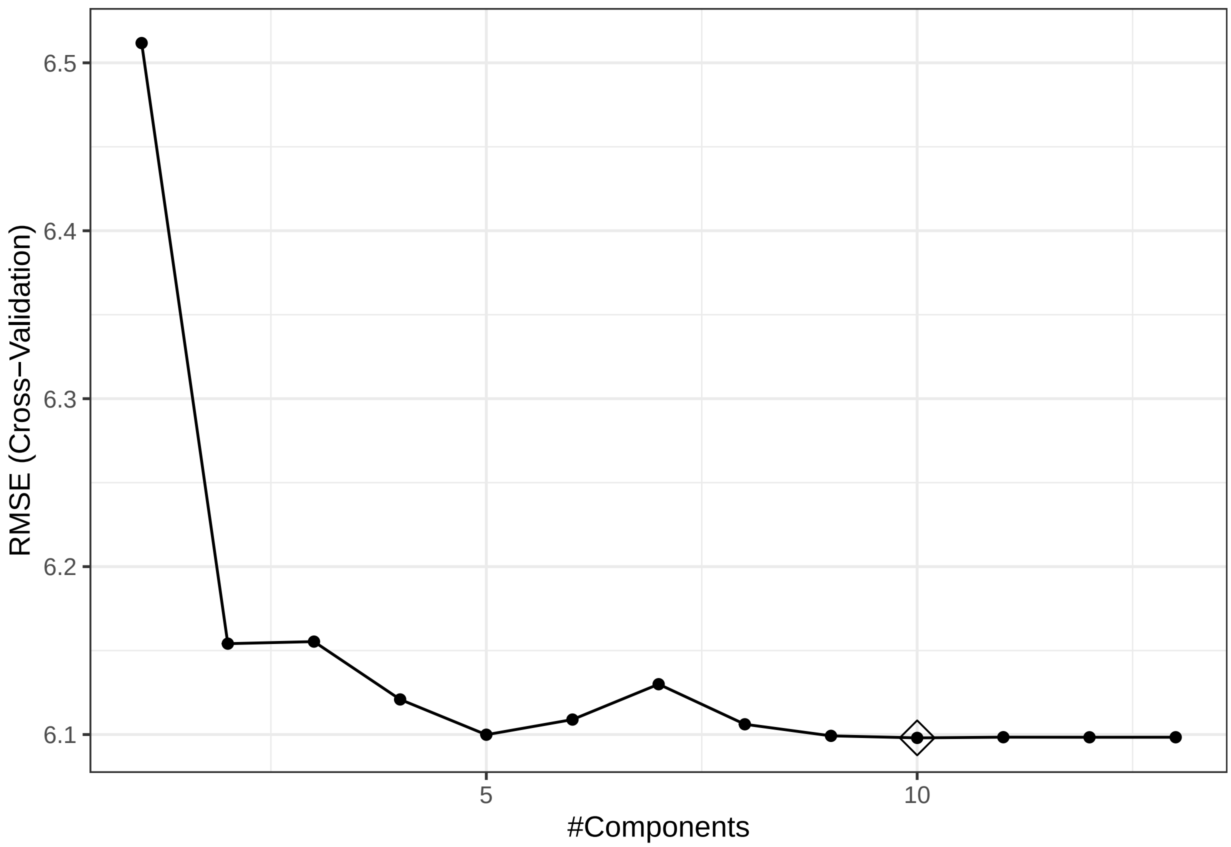

Figure A5. Tuning Optimal Components of PLS Model

MARS Model

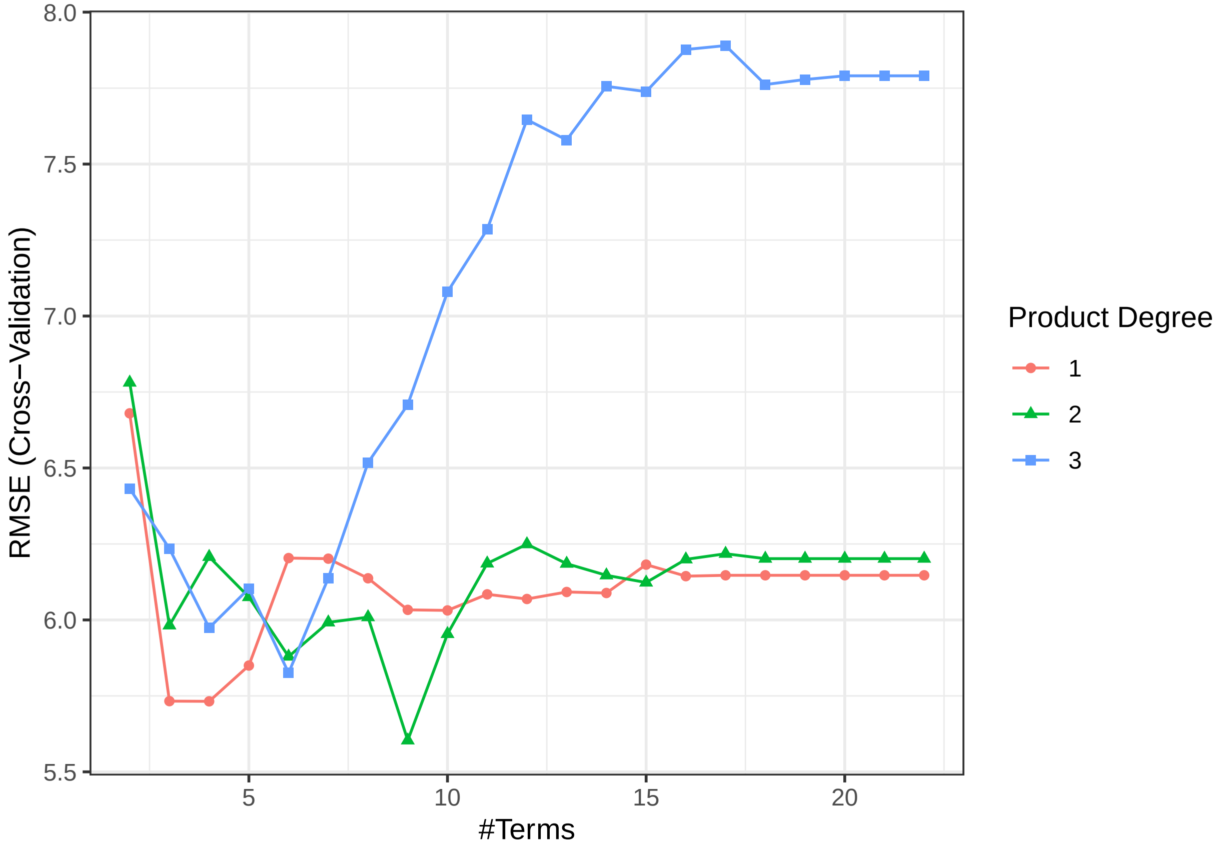

Figure A6. Tuning MARS Model

Regression Tree Model

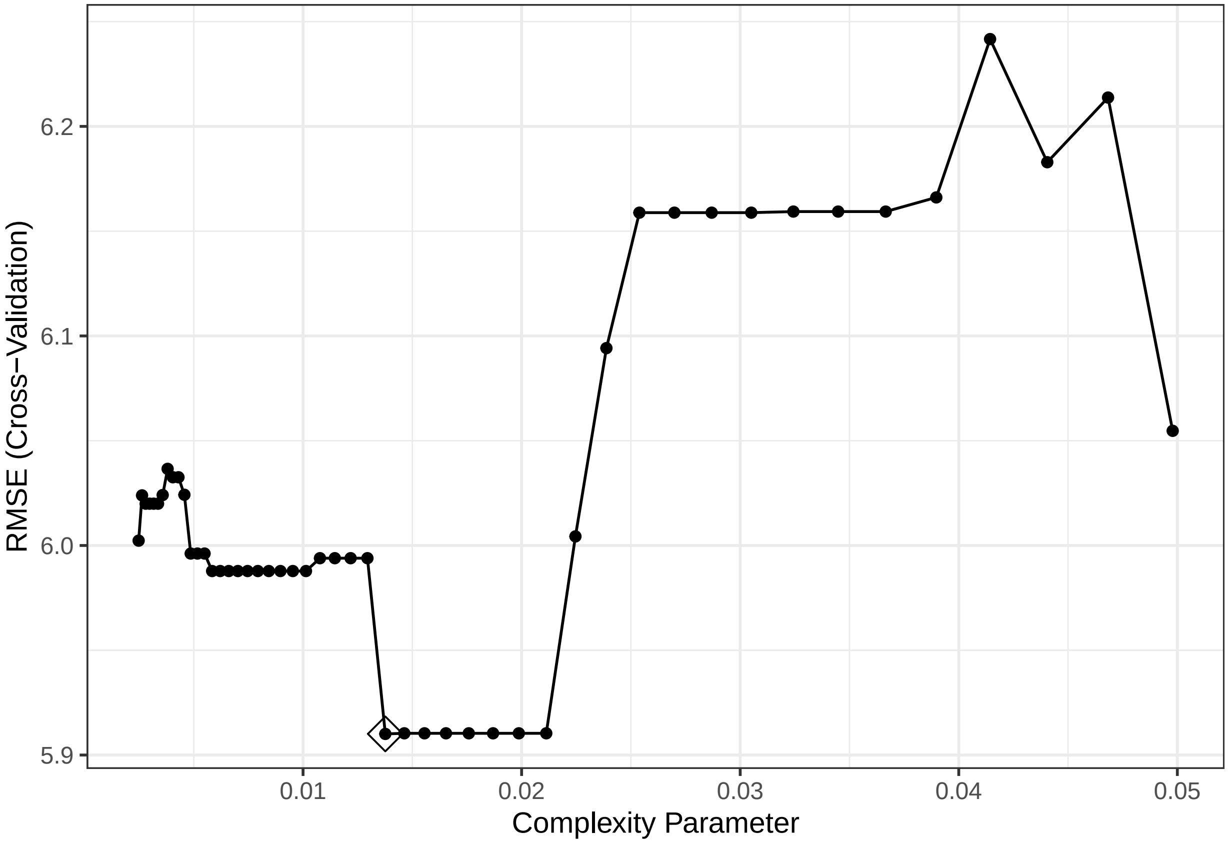

Figure A7. Tuning Complexity Parameter of Regression Tree Model

Conditional Inference Tree Model

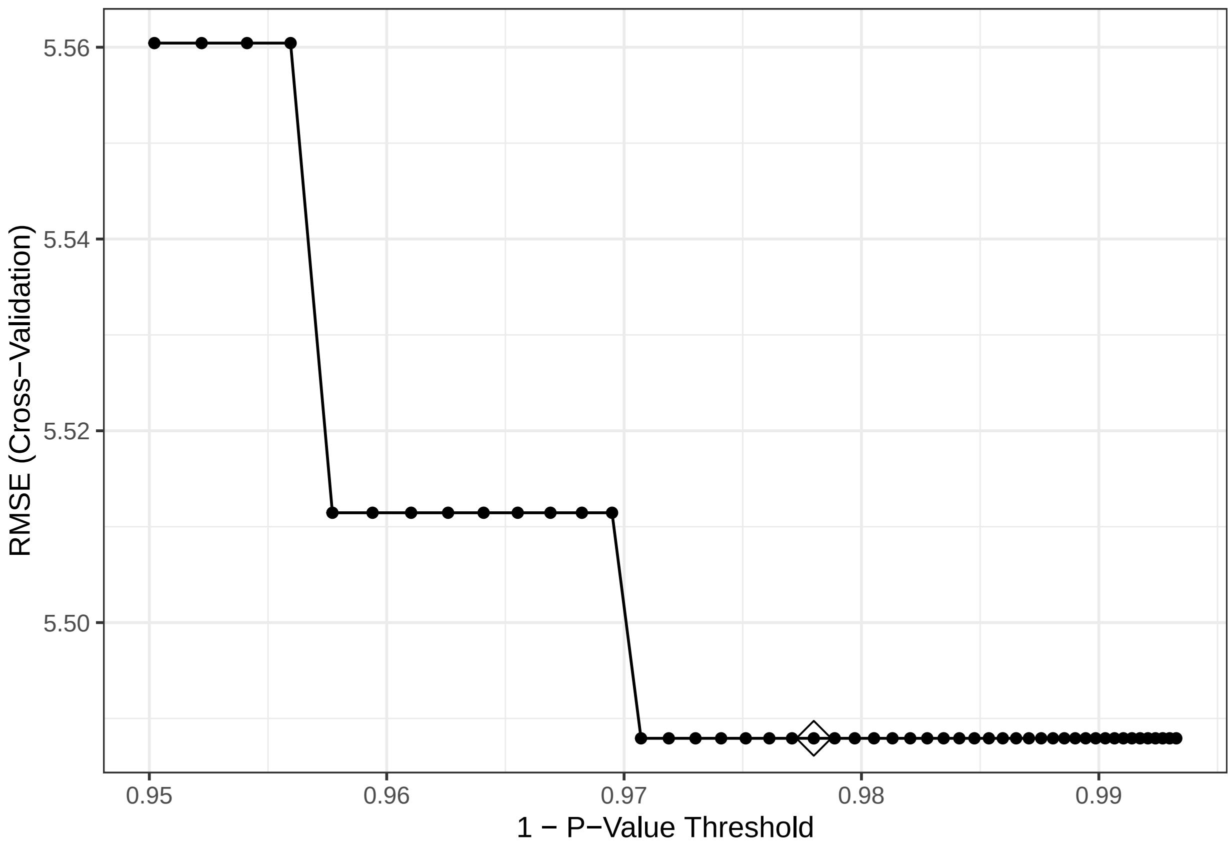

Figure A8. Tuning Mincriterion of Conditional Inference Tree Model

Random Forest Model

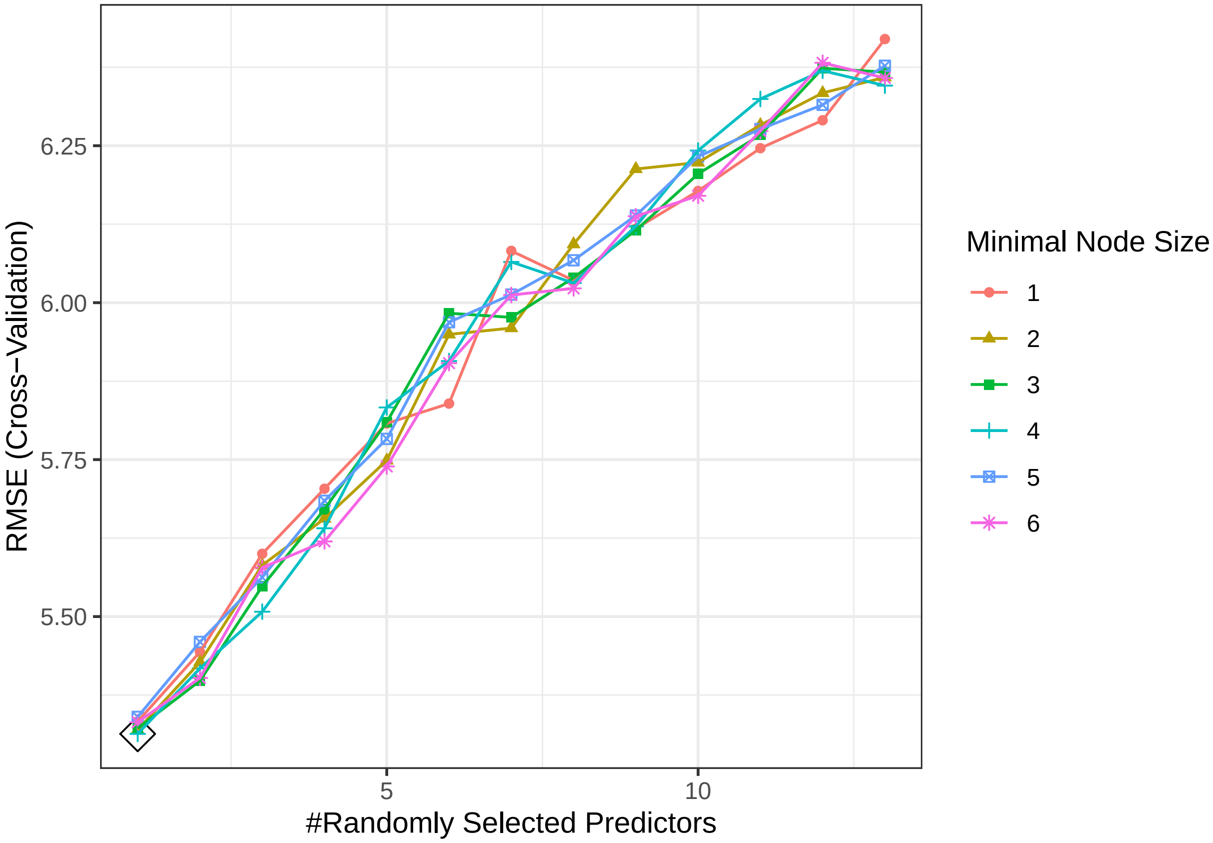

Figure A9. Tuning Random Forest Model

Boosting Model

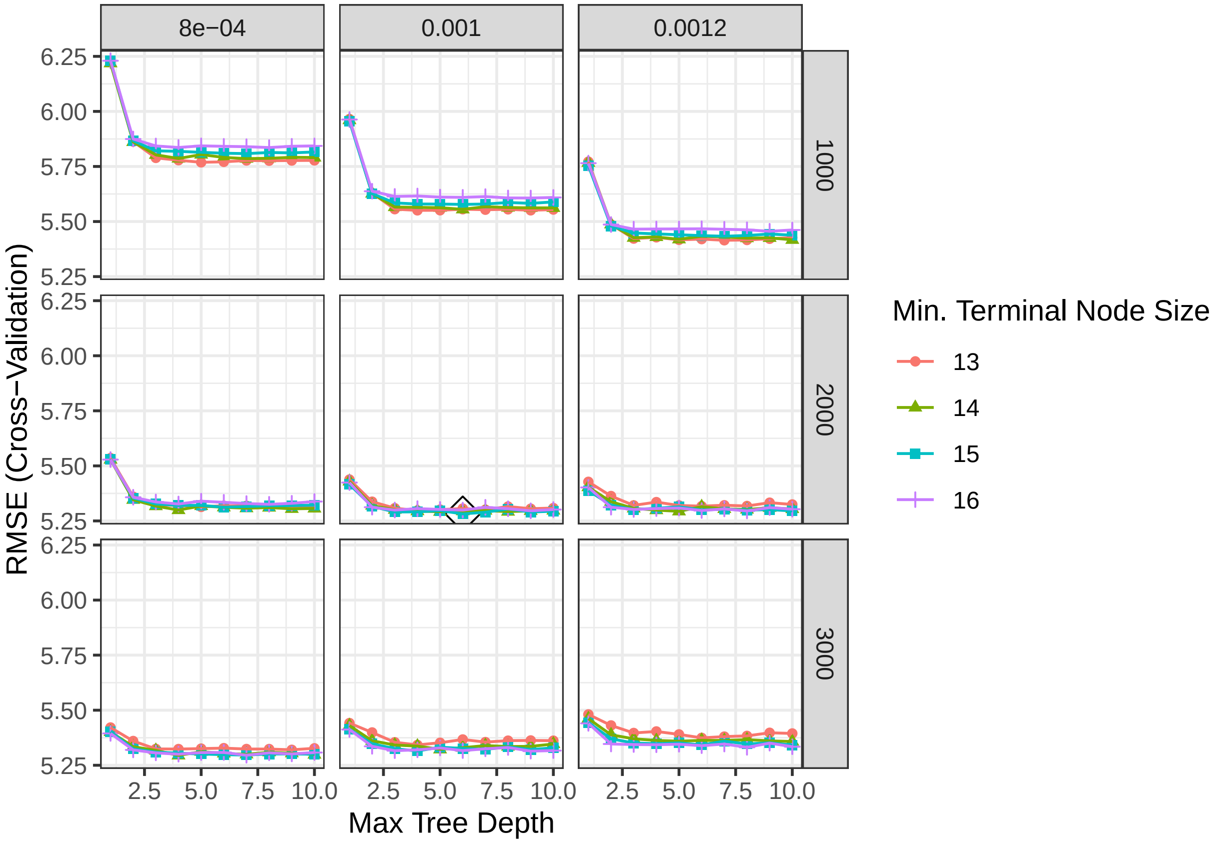

Figure A10. Tuning Boosting Model