Predicting Crime Rate Within Boston Area

Zhourong Li (zl2977), Cong Zhang (cz2612), Ruiyang Li (rl3034)

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 a small proportion of missing values, we may consider using r functions such as `na.omit` or `drop_na()` to remove them; otherwise, imputation may be needed. 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 variables. 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, the dummy variable indicating whether the tract of the Charles River bounds river.

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 regression, ridge regression, 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 the models that require tuning parameters, we specified a grid of numbers and then selected the one that gave the smallest cross-validation RMSE value. 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, which are ZN, DIS, RAD, B, and MEDV.

(2) Ridge Regression: We tuned 100 lambda values in between e^2 to e^{-3}. From 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 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 Figure A3 in the Appendix we can see that 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. 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 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 used all 13 predictors and takes the form of crime rate = CHASotherwise + RAD + s(ZN) + s(PTRATIO) + s(TAX) + s(INDUS) + s(NOX) + s(MEDV) + s(B) + s(AGE) + s(DIS) + s(RM) + s(LSTAT). Specifically, ZN, PTRATIO, TAX, INDUS, NOX, and AGE have a linear association with crime rate whereas the rest of the predictors have a non-linear association.

(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 = 0.014 (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 = 0.98 (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 1 selected variable at each split and 4 minimum node size (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 (black hollow point in Figure 4) 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 not sensitive 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-linear association 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 a black-box model, the random forest model does not have an explicit form. However, it can be interpreted, both globally and locally, using visualization tools.

Based on the variable importance plot (VIP) in Figure 5A, MEDV, DIS, RAD, TAX, and B are the top 5 variables of importance in predicting the crime rate. The individual conditional expectation (ICE) curves (Figure 5B) illustrate how these variables impact the prediction of the response variable. For example, the crime rate seems to mildly fluctuate when the median value of owner-occupied homes (MEDV) is below 10 thousand dollars. It decreases quickly when the median value is between 10 and 12 thousand dollars, and decreases slowly when the median value is greater than 12 thousand dollars. When the median value exceeds 17 thousand dollars, the crime rate becomes stable. Similarly, we can see that the crime rate seems to first decrease with weighted distances to five Boston employment centers (DIS) and then remain stable after the weighted distances is around 1.9. Overall, the per capita crime rate seems to be stable when the index of accessibility to radial highways (RAD) is below 14. It starts to slowly increase after the index exceeds 14, and quickly jumps up when the index is around 16. However, when the index is greater than 16, the per capita crime rate seems to become stable again. Furthermore, we can see that the crime rate seems to be relatively stable when the full-value property-tax rate per $10,000 (TAX) is below 540, and it increases when TAX is between 540 and 570. After remaining stable when TAX is between 570 and 680, the crime rate decreases a little and then becomes stable again when TAX exceeds 690. In addition, the per capita crime rate seems to quickly decrease when the index of proportion of blacks by town (B) is below 20, and it slowly decreases when the index B is greater than 20.

We can also visualize the result of our final model on new observations, to understand our final model more and to evaluate its performance. We used the first 20 test data observations for illustration. Based on the LIME plot (Figure 5C), we can see that the feature weight tends to be positive if the median value of owner-occupied homes (MEDV) is below its first quartile (Q1) 16.9 thousand dollars, but tends to be negative if MEDV is above 16.9 thousand dollars. This implies that lower MEDV is associated with higher crime rate. Similar pattern can also be observed for the index B, the average number of rooms (RM), and the proportion of residential land (ZN) – lower B (meaning proportion of blacks is around 0.63) or lower RM or lower ZN is associated with higher crime rate, though the effects seem to be weaker. On the contrary, the feature weight of the index of accessibility to radial highways (RAD) tends to be negative if RAD is below 5, but tends to be positive if RAD is above 5; this indicates that higher RAD is associated with higher crime rate. Similarly, we can see that higher full-value property-tax rate (TAX), or higher pupil-teacher ratio (PTRATIO), or higher percentage of lower status of the population (LSTAT) is associated with higher crime rate. All these interpretations are reasonable; therefore, we considered our final model to be reliable.

Conclusion and 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.

Overall, our model suggests that the median value of owner-occupied homes, distances to employment centers, and accessibility to radial highways are the top 3 variables that play important roles in predicting the crime rate. Specifically, the crime rate is predicted to decrease with higher median home value, to decrease if the distance to employment centers is farther, and to increase with the accessibility to radial highways when the index of latter exceeds 14. These findings are consistent with our natural understanding and expectation and provide meaningful insight into public safety. For example, communities with higher median home values may have better security facilities, which leads to lower crime rate; longer distance to employment centers may discourage the criminals of those centers from traveling to the communities; better accessibility to radial highways would allow the criminals to flee more easily, which could be a potential incentive for crimes. Therefore, the policymakers or authorities may consider making more efforts 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.

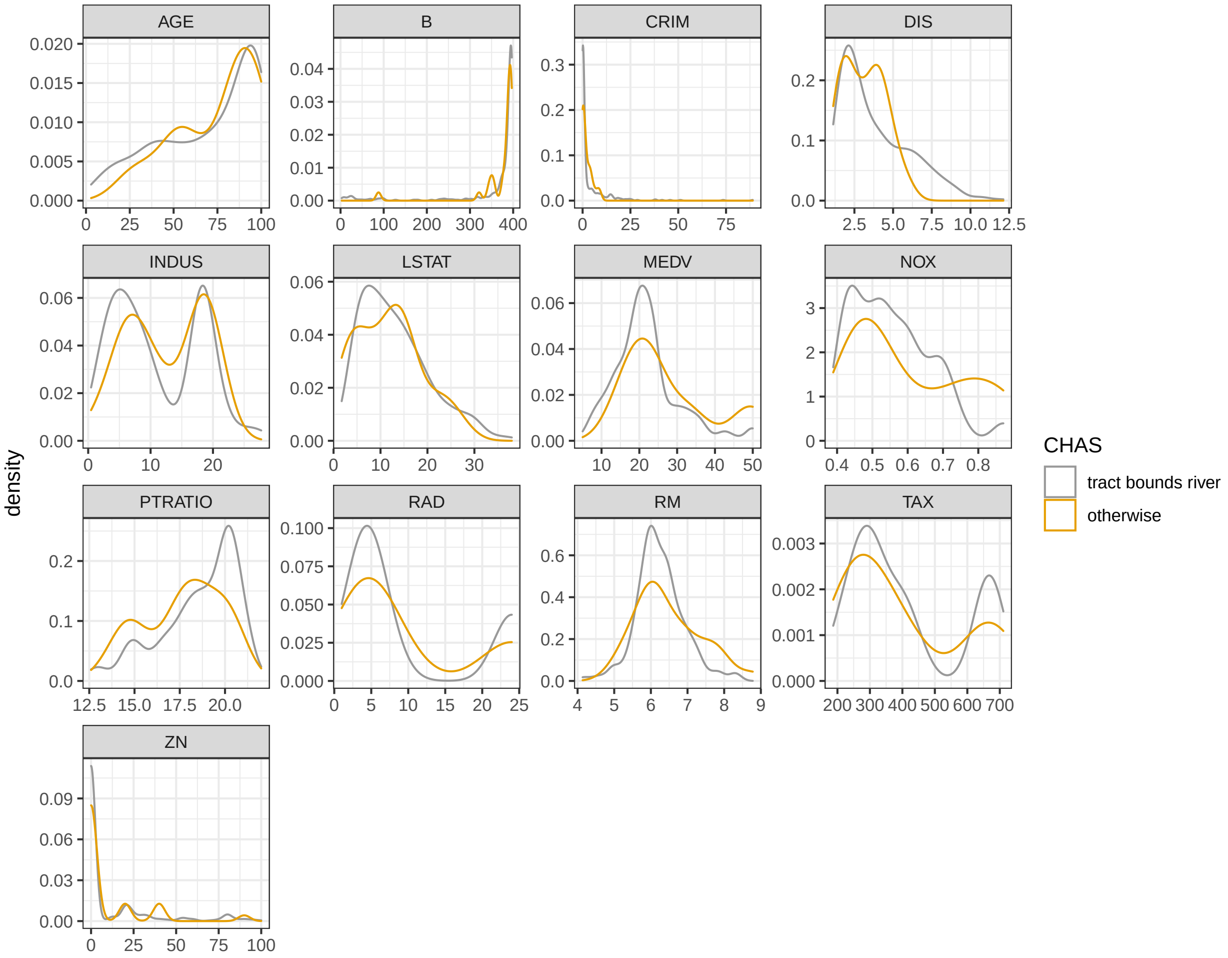

Figure 1. Variable Distribution (Training Dataset)

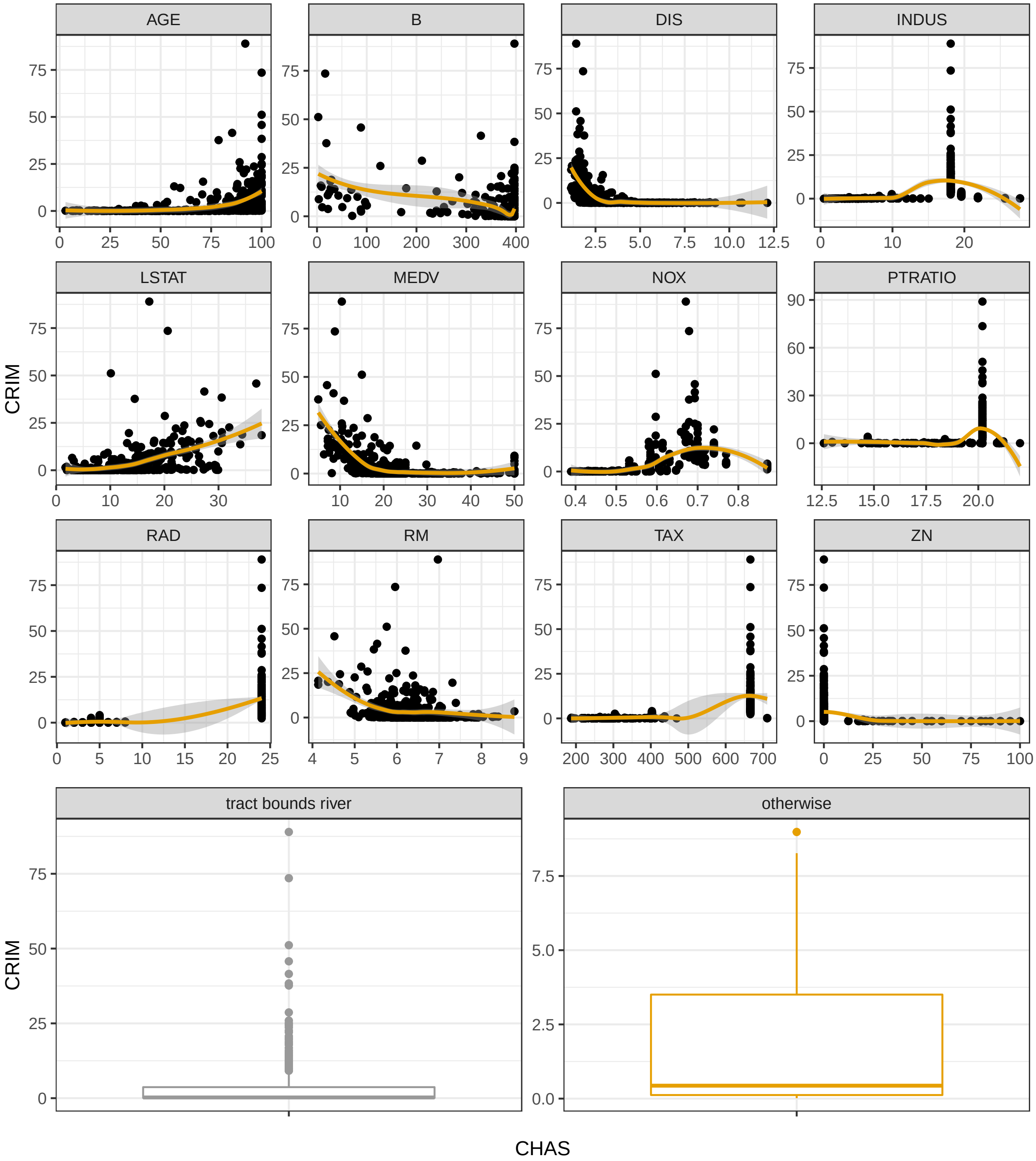

Figure 2. Variables vs. Crime Rate (Training Dataset)

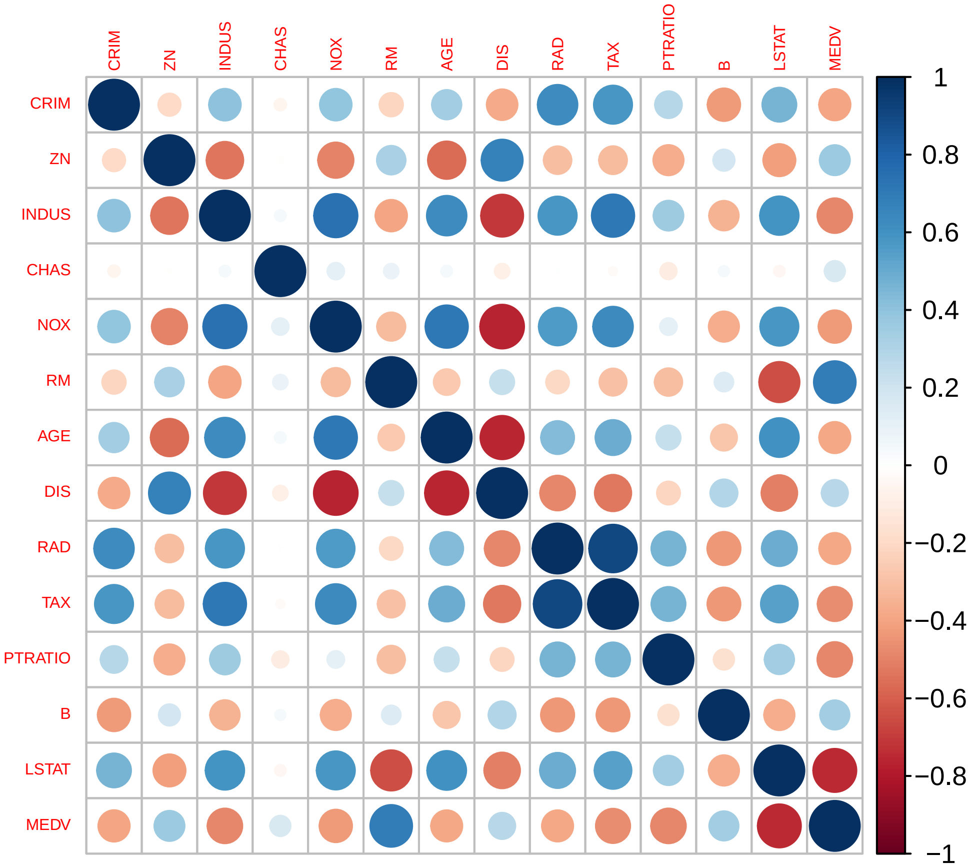

Figure 3. Correlation Plot (Training Dataset)

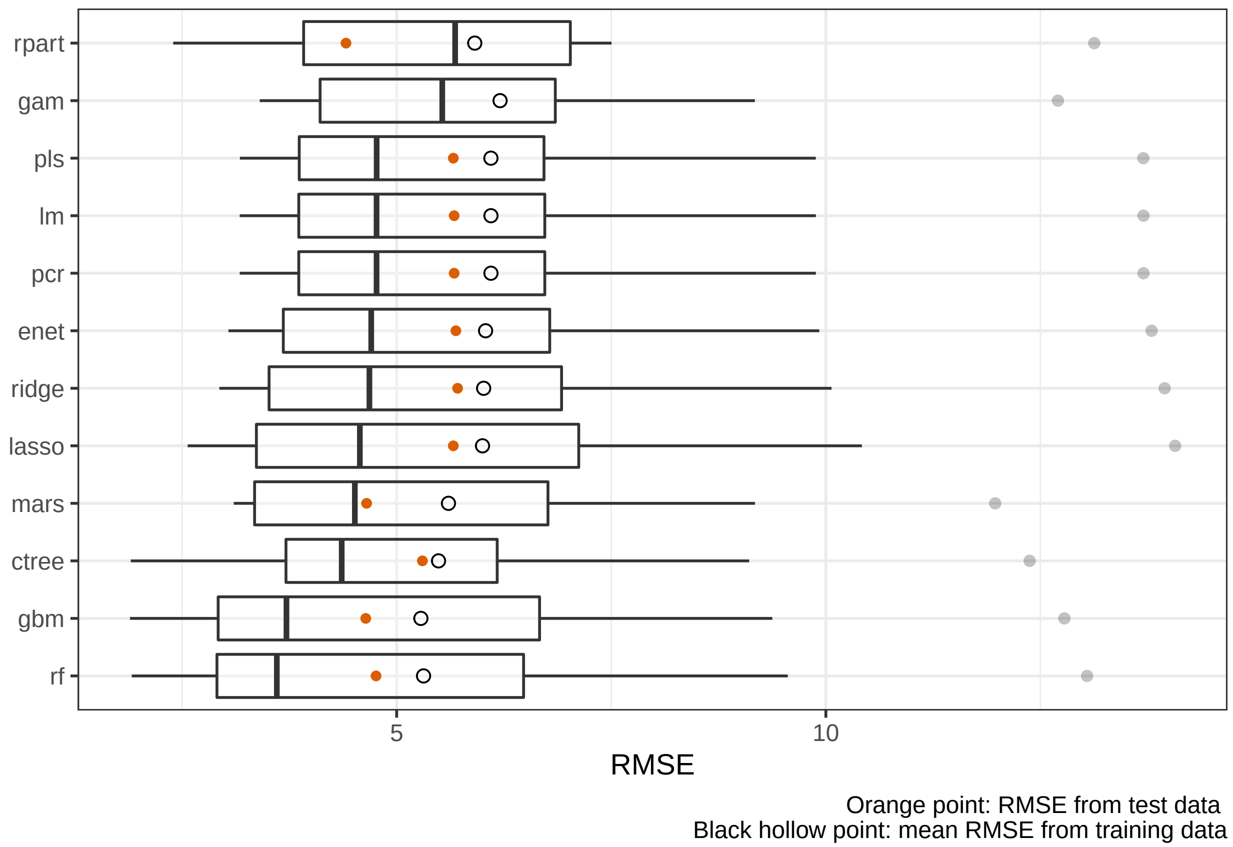

Figure 4. Predictive Performance of All Models

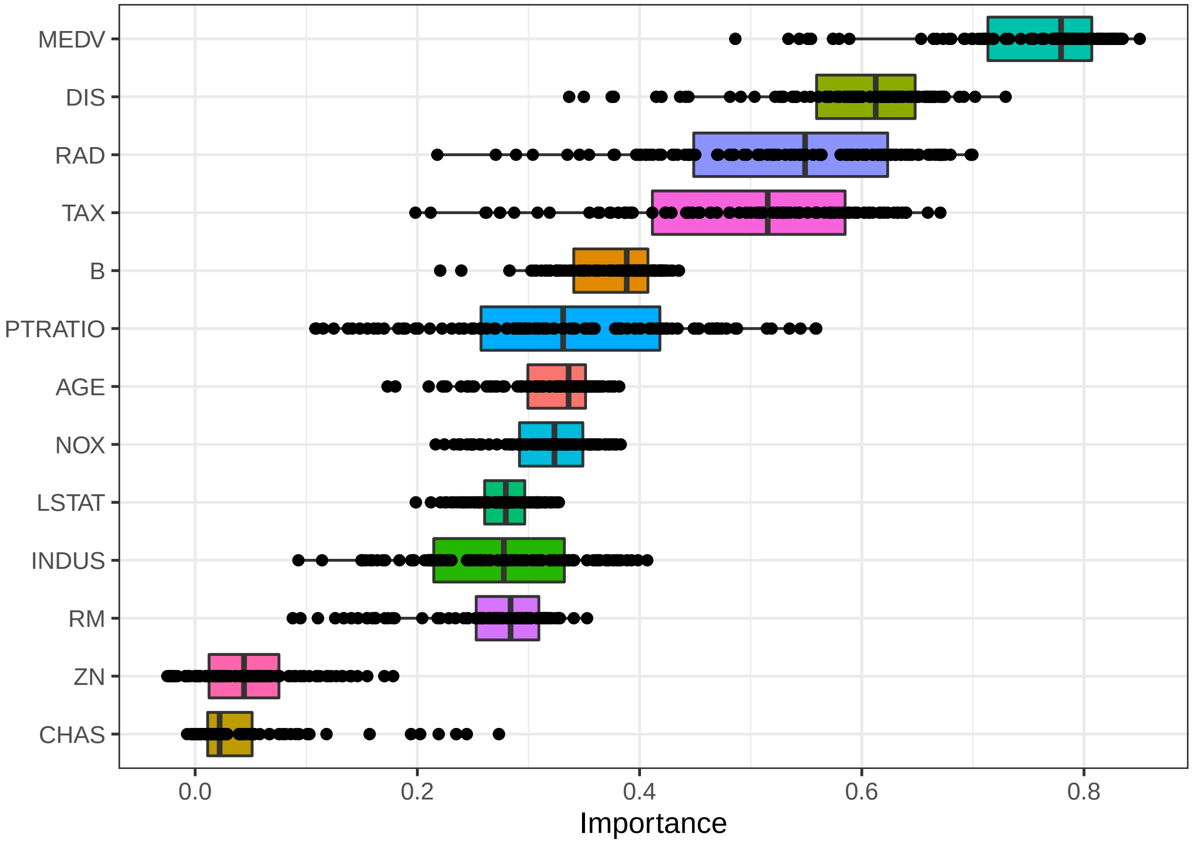

Figure 5A. The Variable Importance Plot (VIP) from Random Forest Model (Final Model)

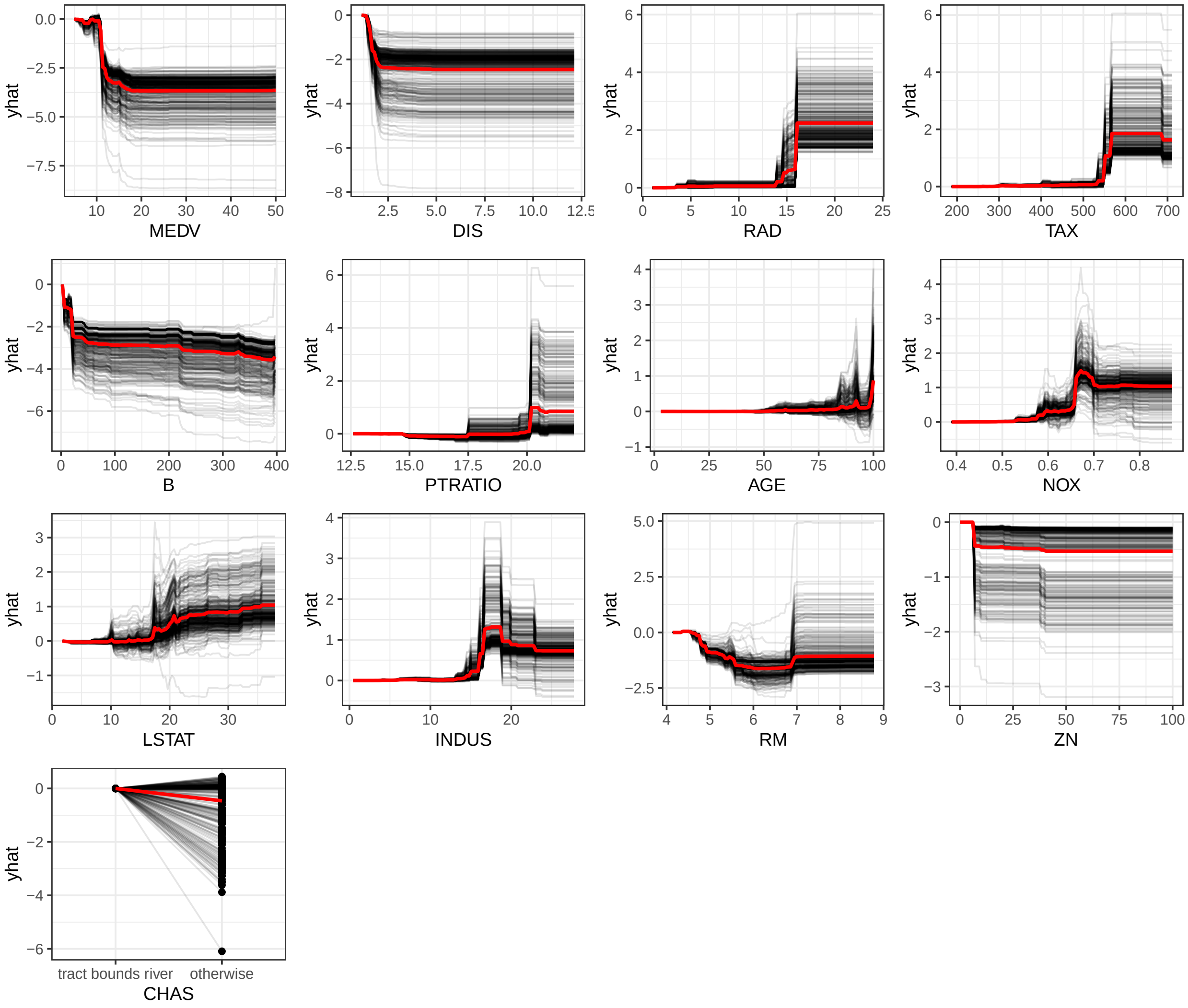

Figure 5B. The Individual Conditional Expectation (ICE) Curves from Random Forest Model (Final Model)

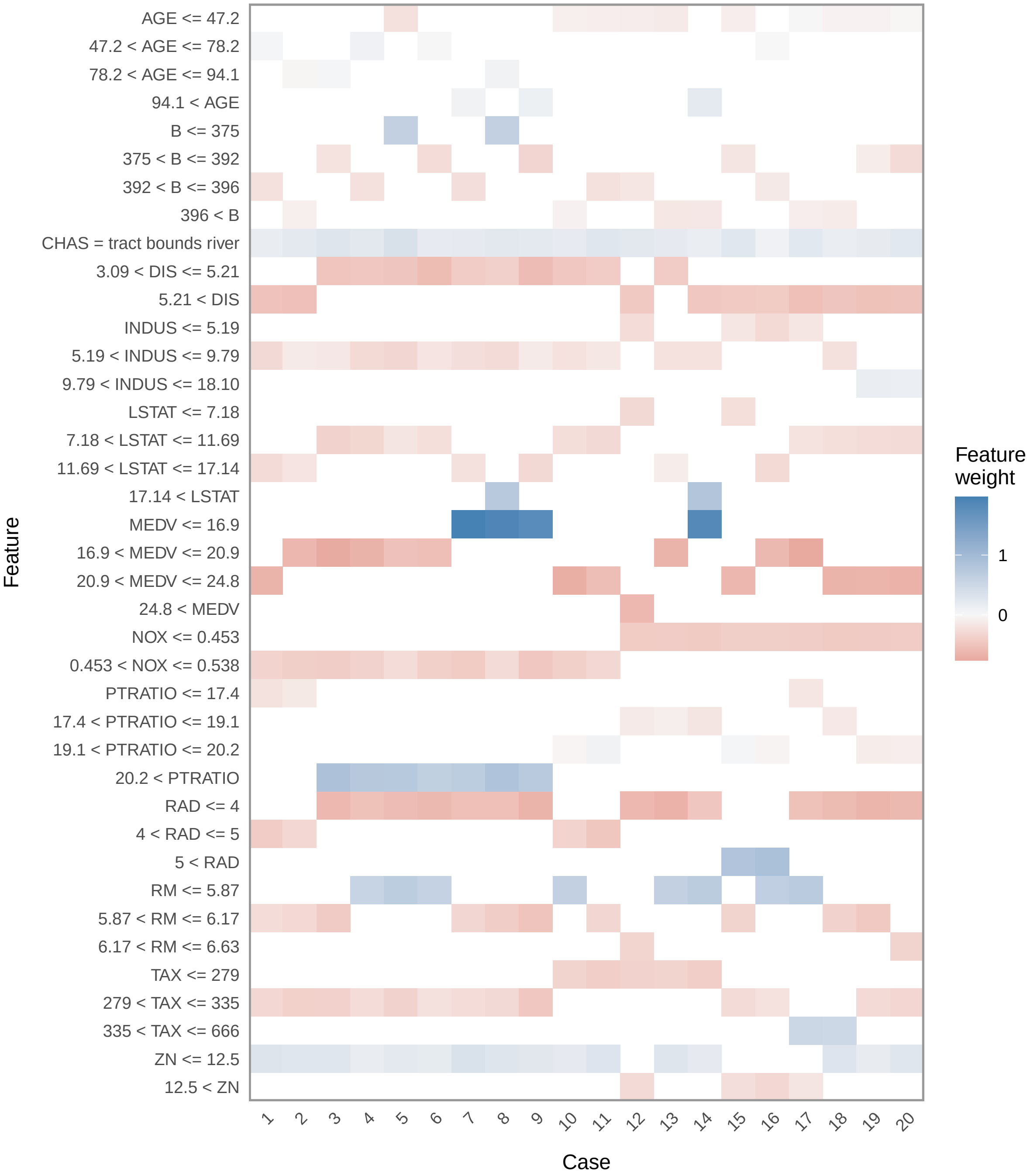

Figure 5C. The Local Interpretable Model-agnostic Explanations (LIME) Plot from Random Forest Model (Final Model)

Appendix of Model Tuning

Ridge Model

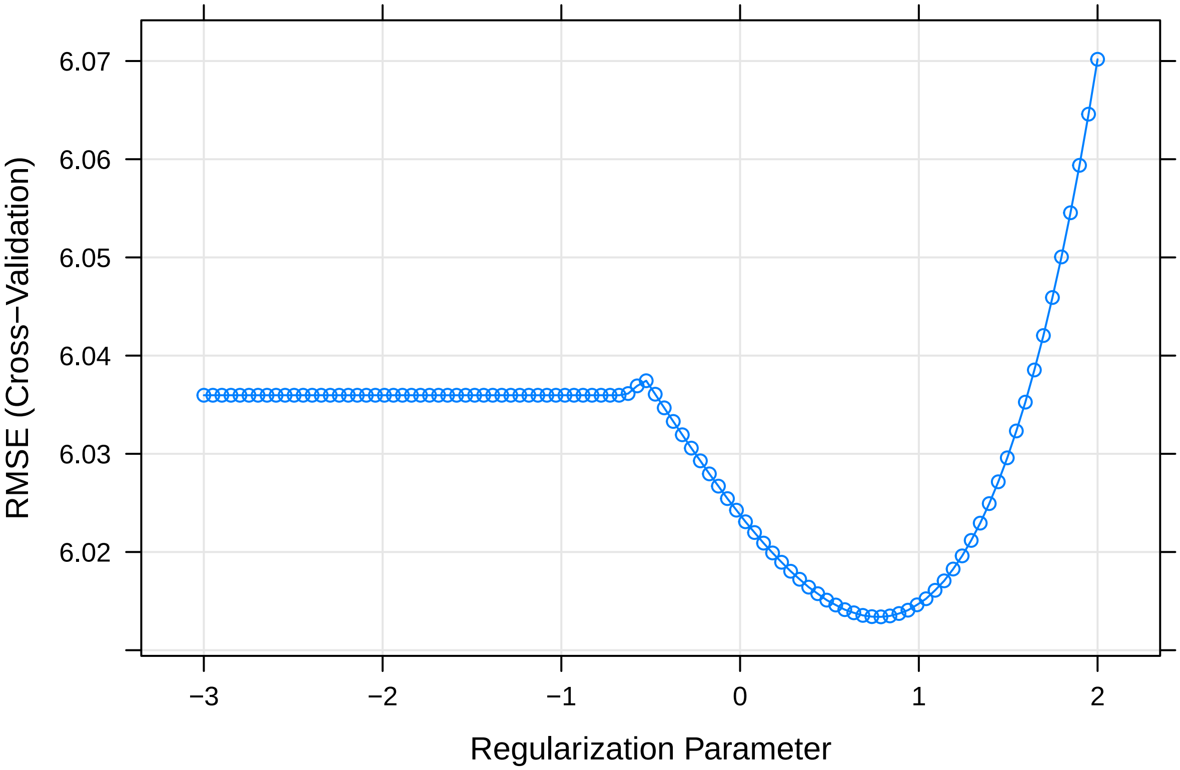

Figure A1. Tuning Lambda of Ridge Model

Lasso Model

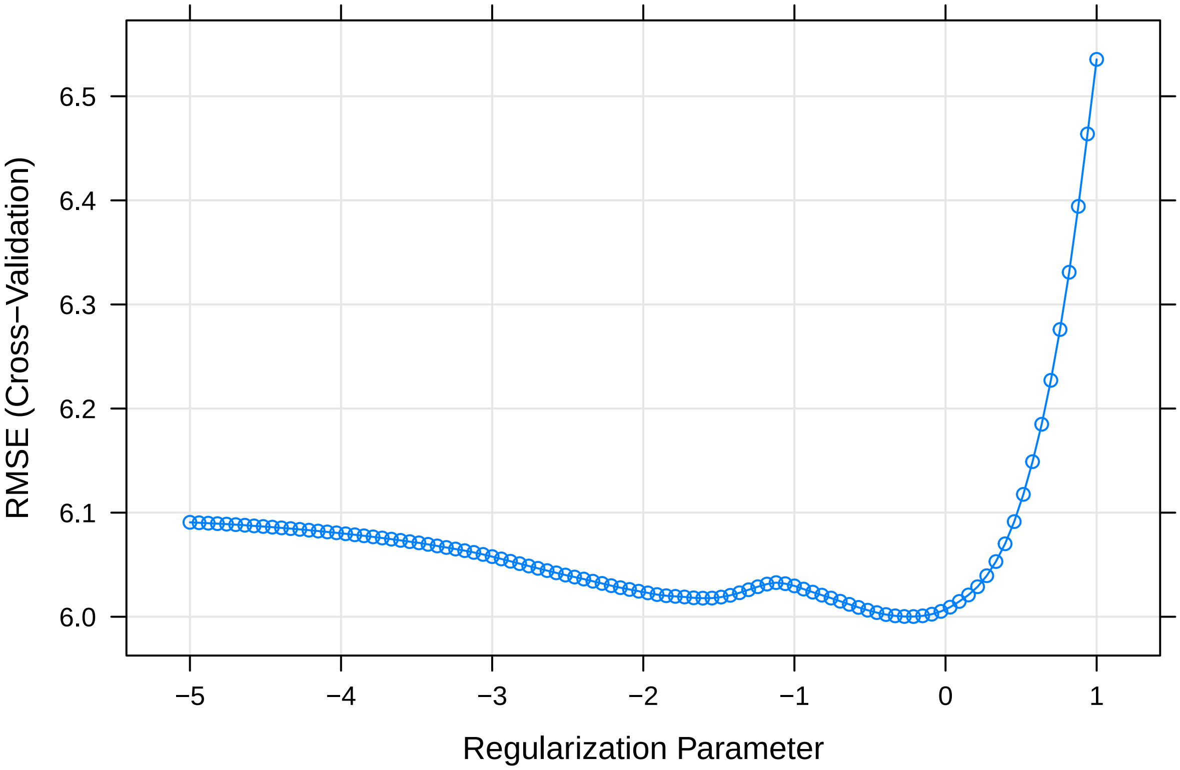

Figure A2. Tuning Lambda of Lasso Model

Elastic Net Model

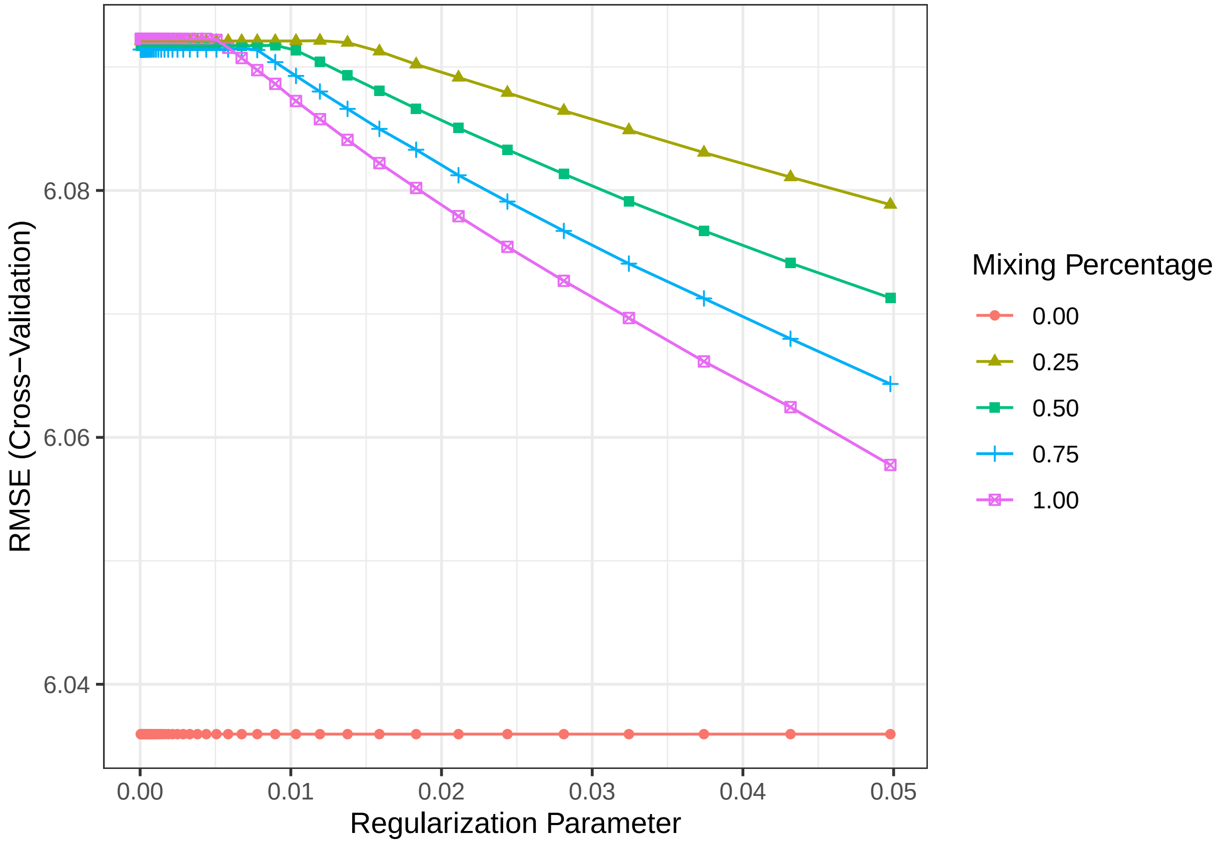

Figure A3. Tuning Alpha and Lambda of Elastic Net Model

PCR Model

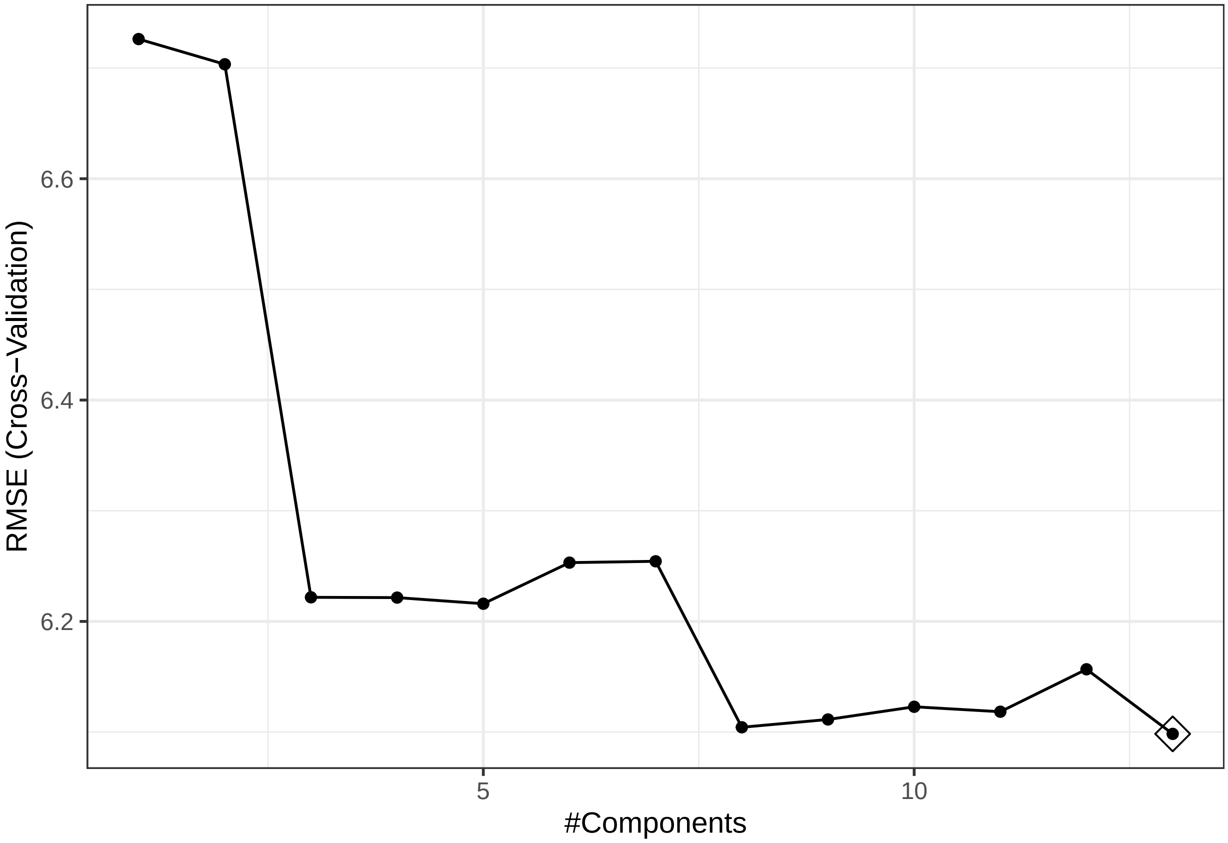

Figure A4. Tuning Optimal Components of PCR Model

PLS Model

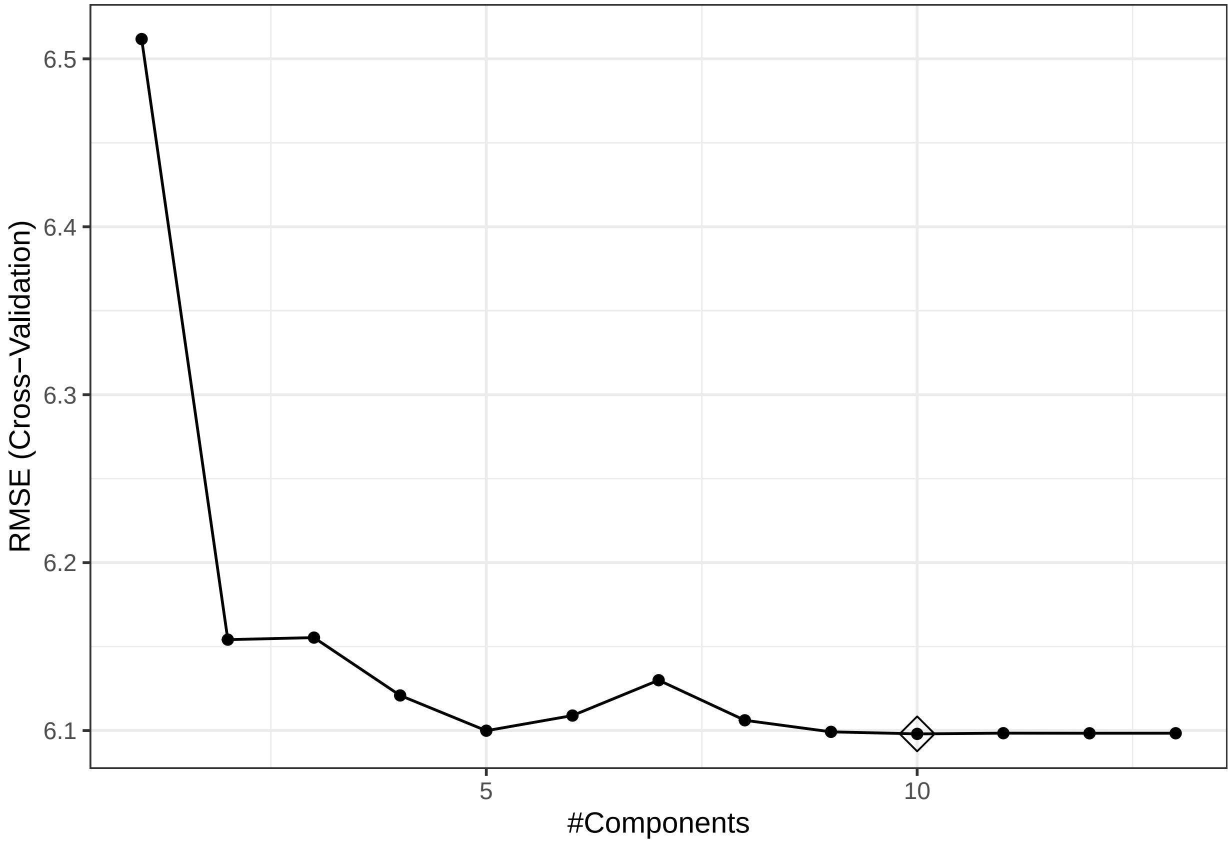

Figure A5. Tuning Optimal Components of PLS Model

MARS Model

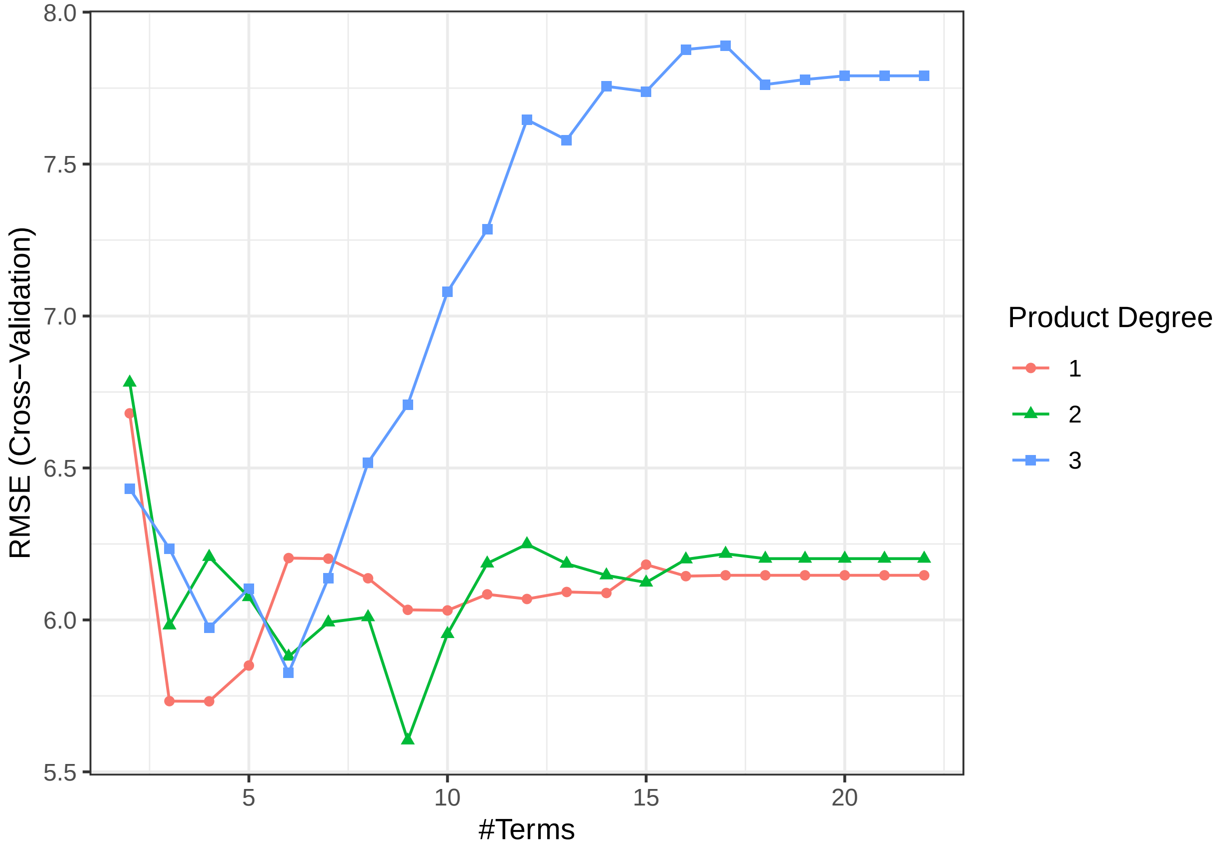

Figure A6. Tuning MARS Model

Regression Tree Model

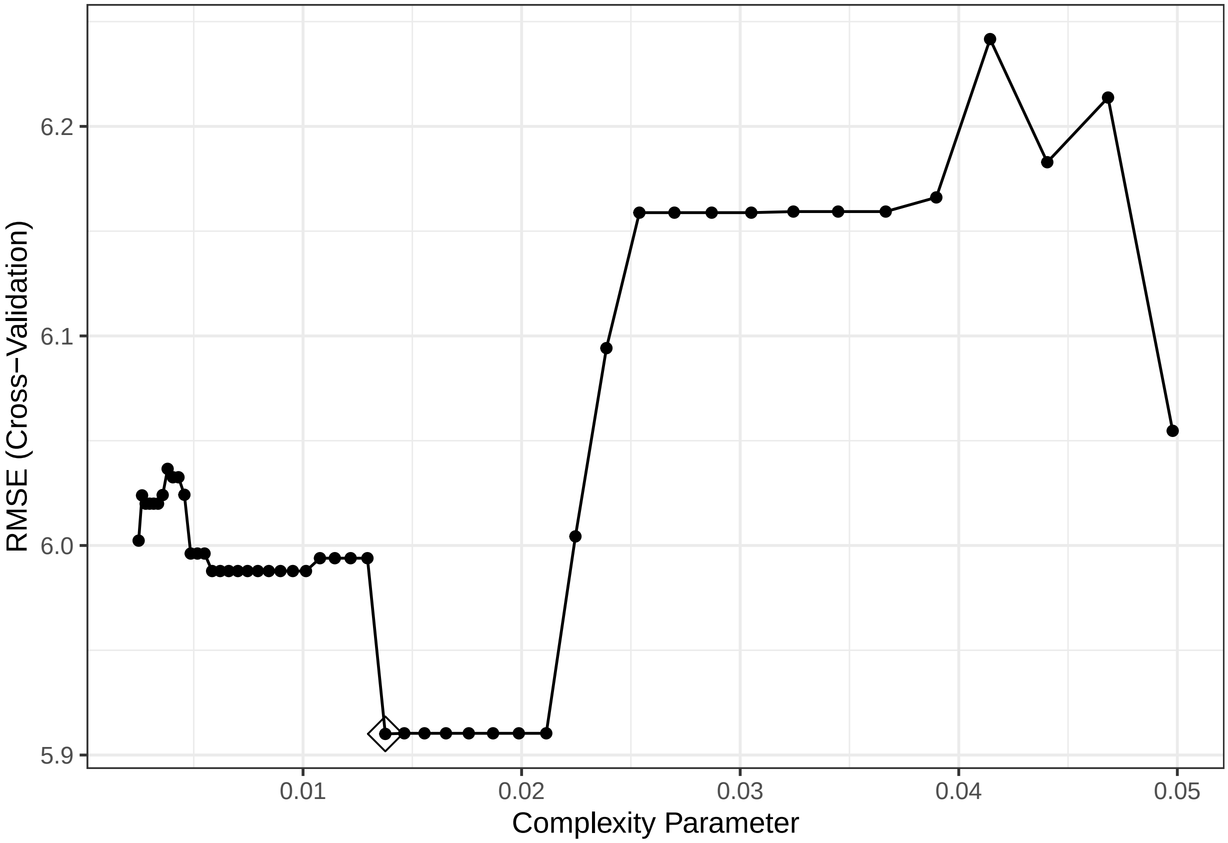

Figure A7. Tuning Complexity Parameter of Regression Tree Model

Conditional Inference Tree Model

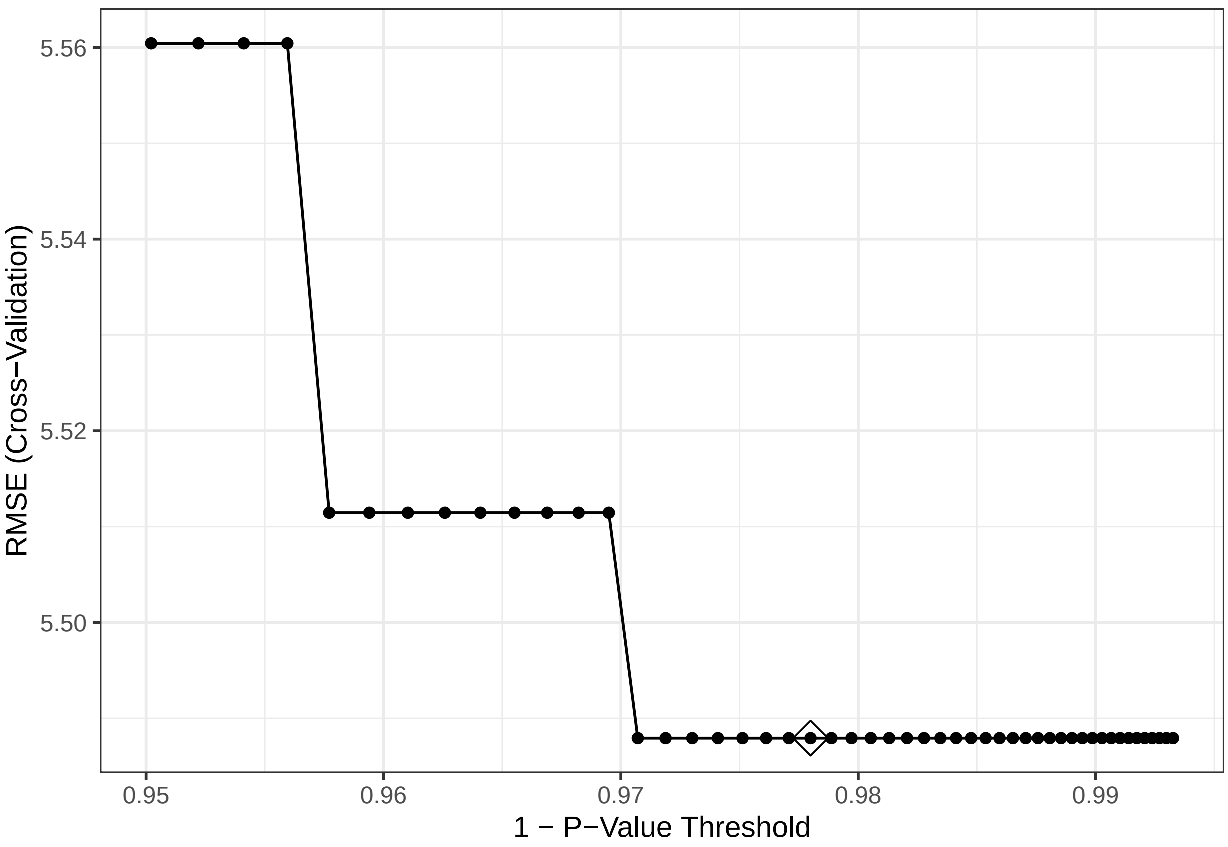

Figure A8. Tuning Mincriterion of Conditional Inference Tree Model

Random Forest Model

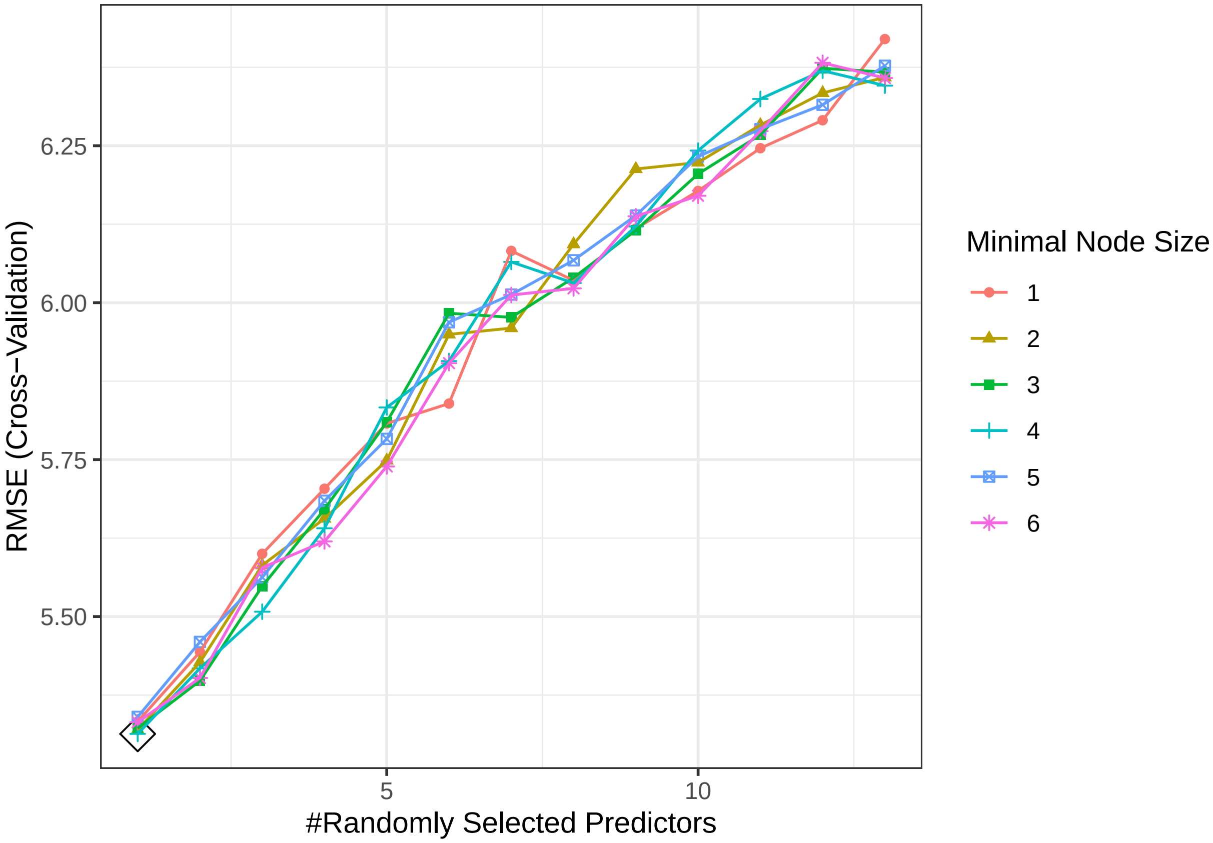

Figure A9. Tuning Random Forest Model

Boosting Model

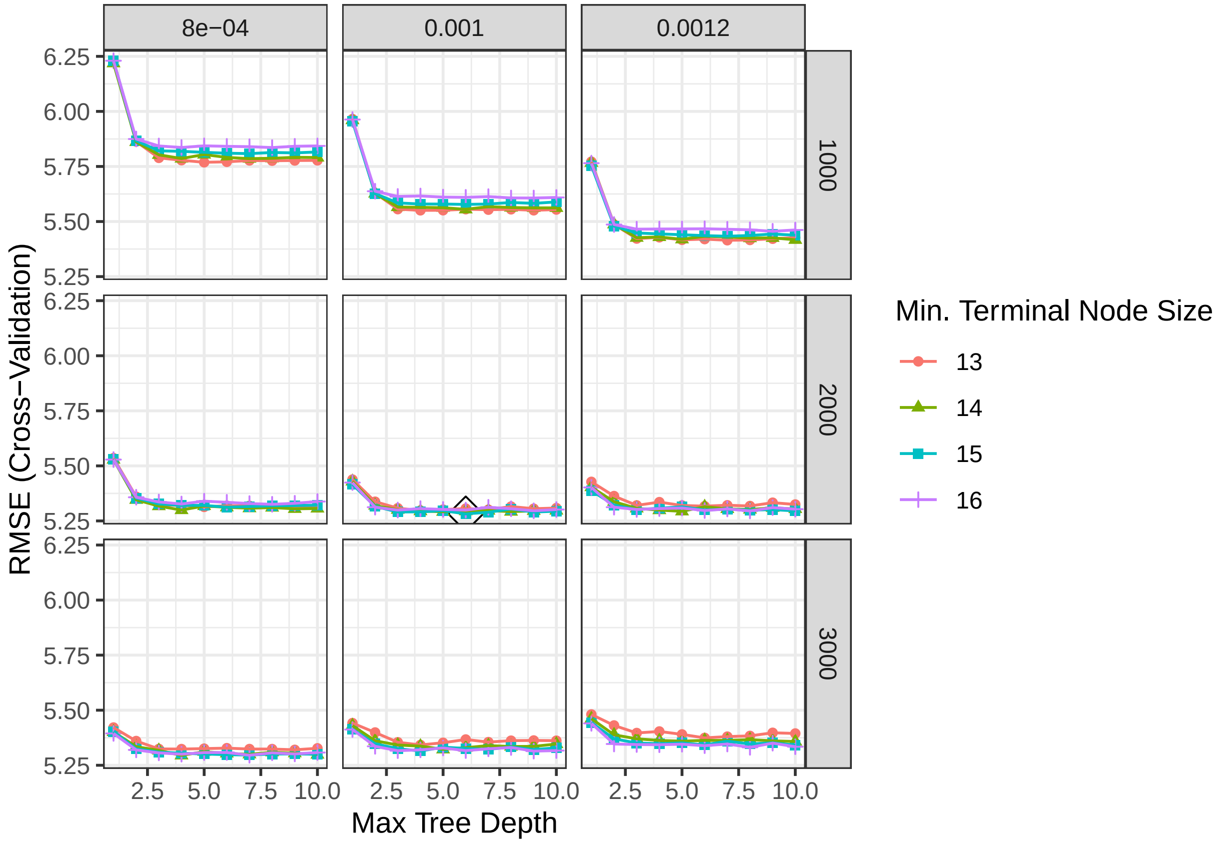

Figure A10. Tuning Boosting Model