Midterm Project Report

Huanyu Chen, Yifei Liu, Longyu Zhang

Abstract

This study examines the predictive performance of various data science models in estimating recovery time from COVID-19 based on demographic and health-related factors. The techniques used include LASSO regression, Elastic Net, PLS, MARS, GAM, and Random Forest. The results show that MARS has the best predictive performance among all the methods. Our findings underscore the importance of considering both linear and nonlinear relationships in modeling recovery time. Moreover, our findings can have implications for clinical decision-making and resource allocation in COVID-19 management.

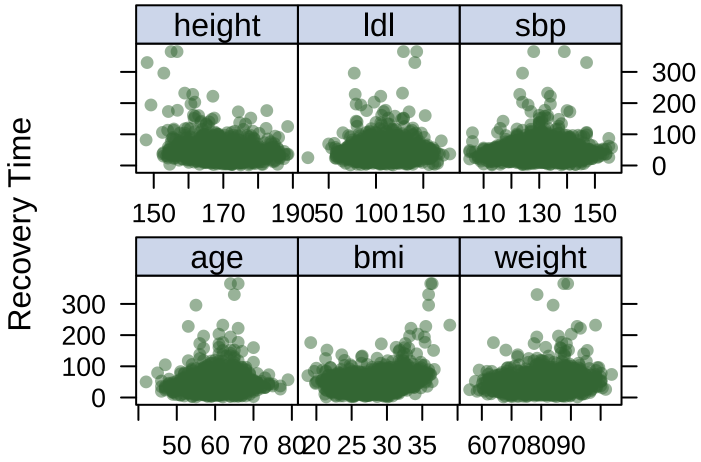
1 Exploratory Analysis and Data Visualization

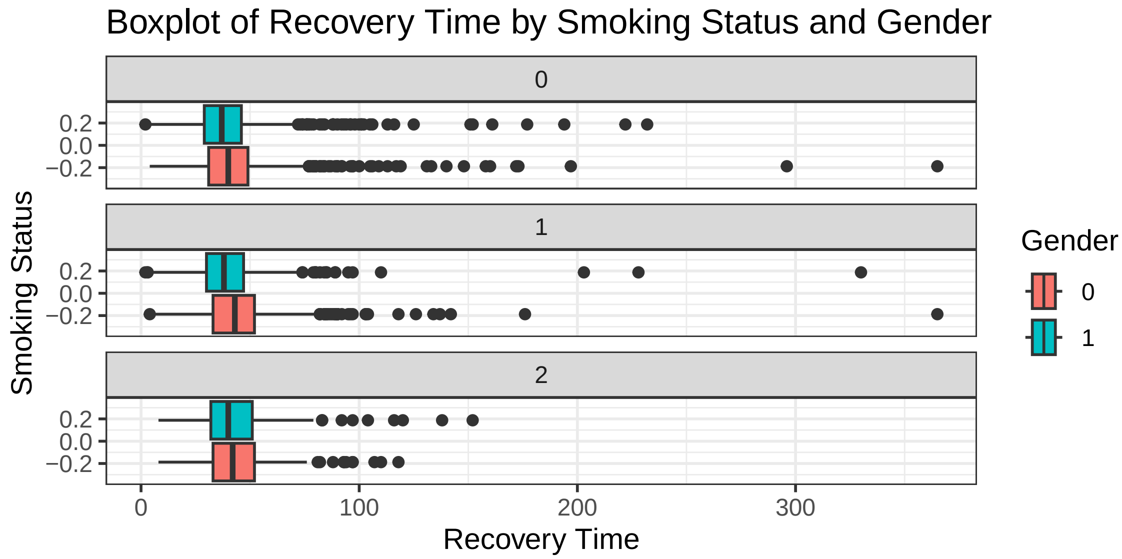
1.1 Variable Types

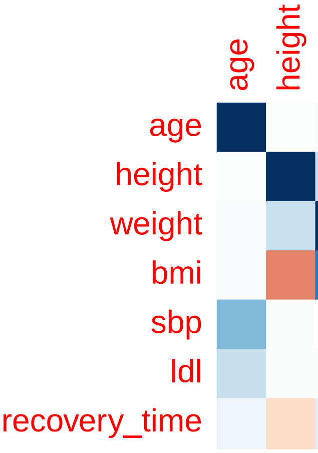
In this dataset, `age`, `height`, `weight`, `bmi`, `SBP`, `LDL`, and `recovery_time` are continuous variables.

1.2 Boxplot of Recovery Time by Smoking Status and Gender

Our analysis reveals a notable trend: across all smoking statuses, females (`gender` = 0) consistently exhibit longer recovery times compared to males. Interestingly, individuals who had never smoked had more outliers on the right side of the boxplot, suggesting a longer recovery time. This counter-intuitive finding suggests that individuals with healthier lifestyles, such as non-smokers, paradoxically require more time to recover from COVID-19.



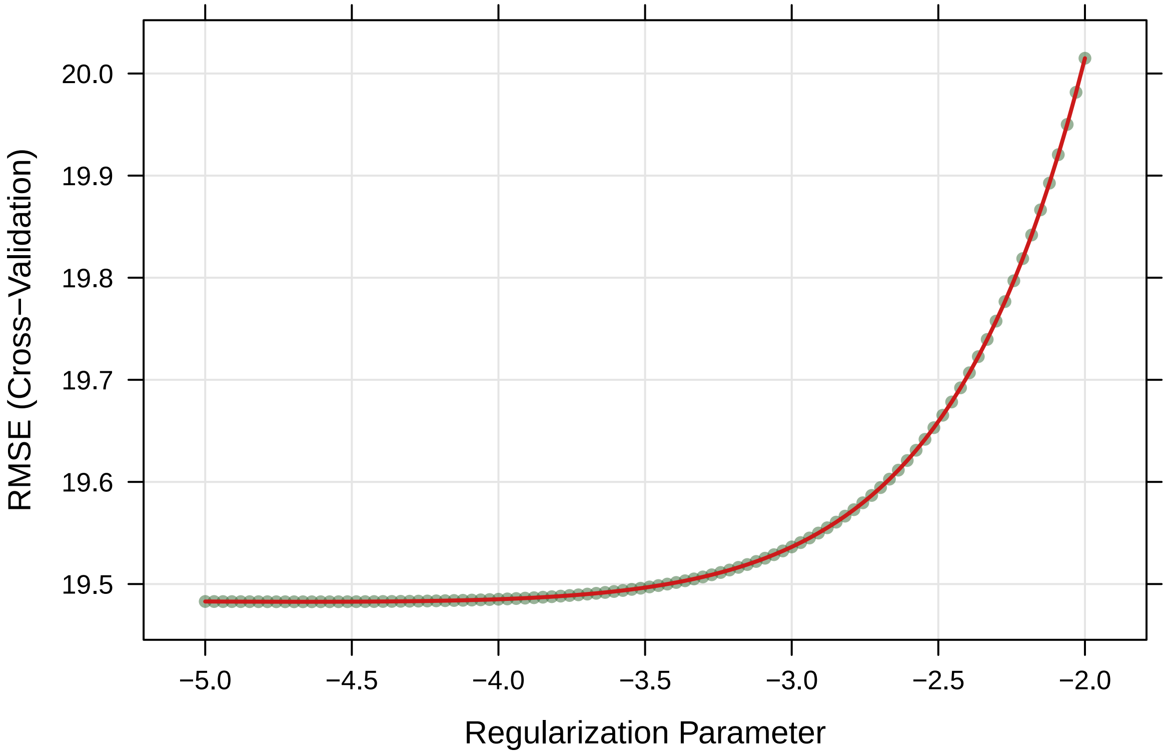


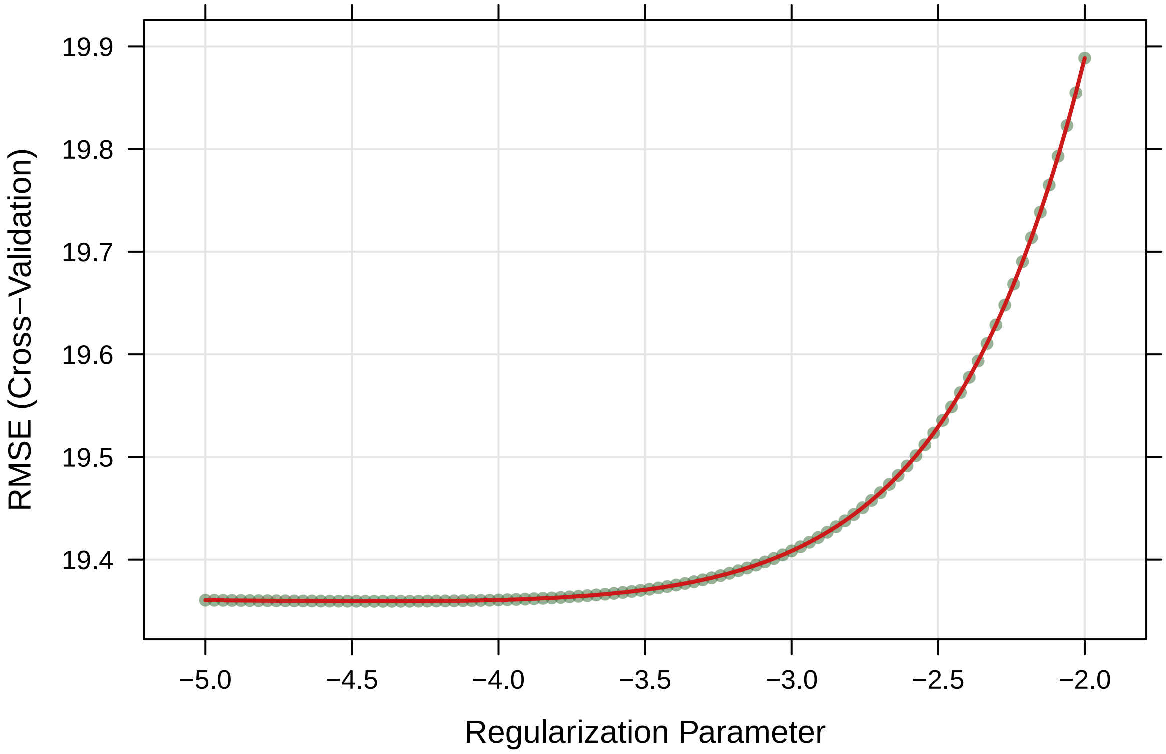


2 Model Training

2.1 Lasso

Model assumptions:
(a) Sparsity Assumption: Lasso assumes that the true model depends on only a small number of predictors, implying that the model is sparse. This means it's suited for scenarios where only a few variables significantly impact the response variable.
(b) Regularization: By penalizing the magnitude of the coefficients (L1 penalty), Lasso encourages smaller absolute values of coefficients, thus reducing model complexity and the risk of overfitting.

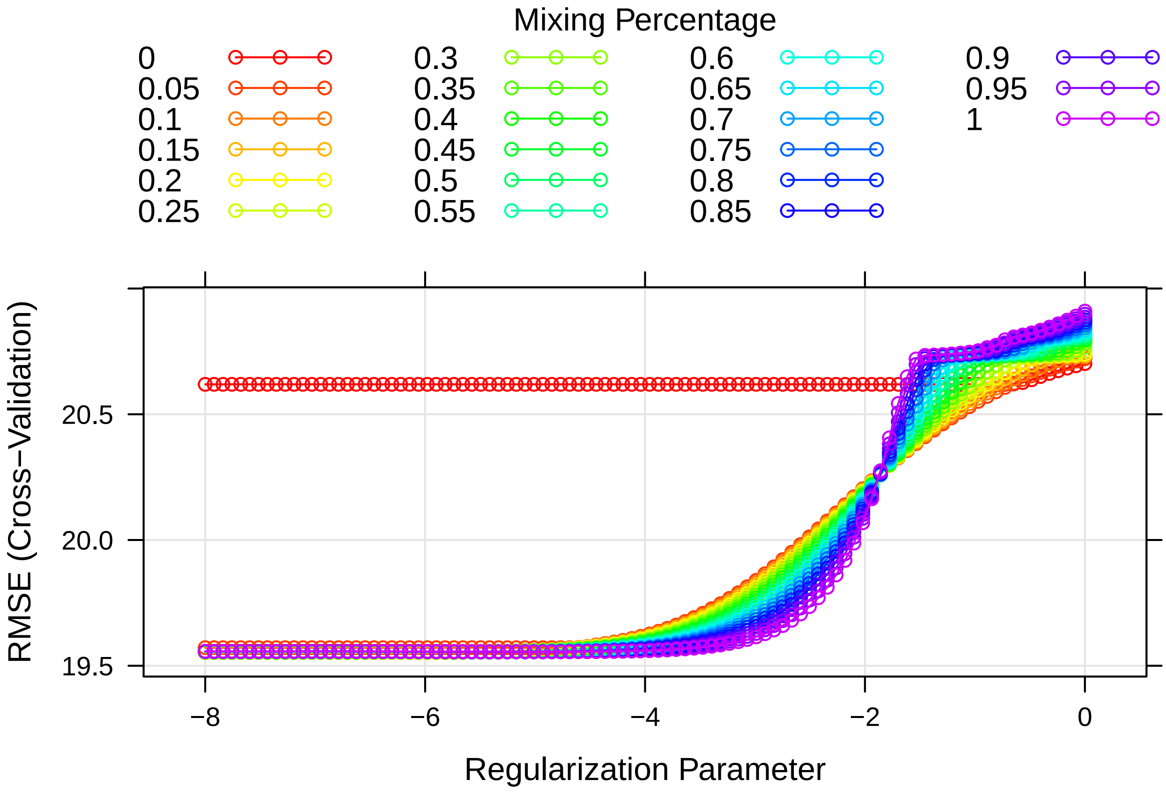


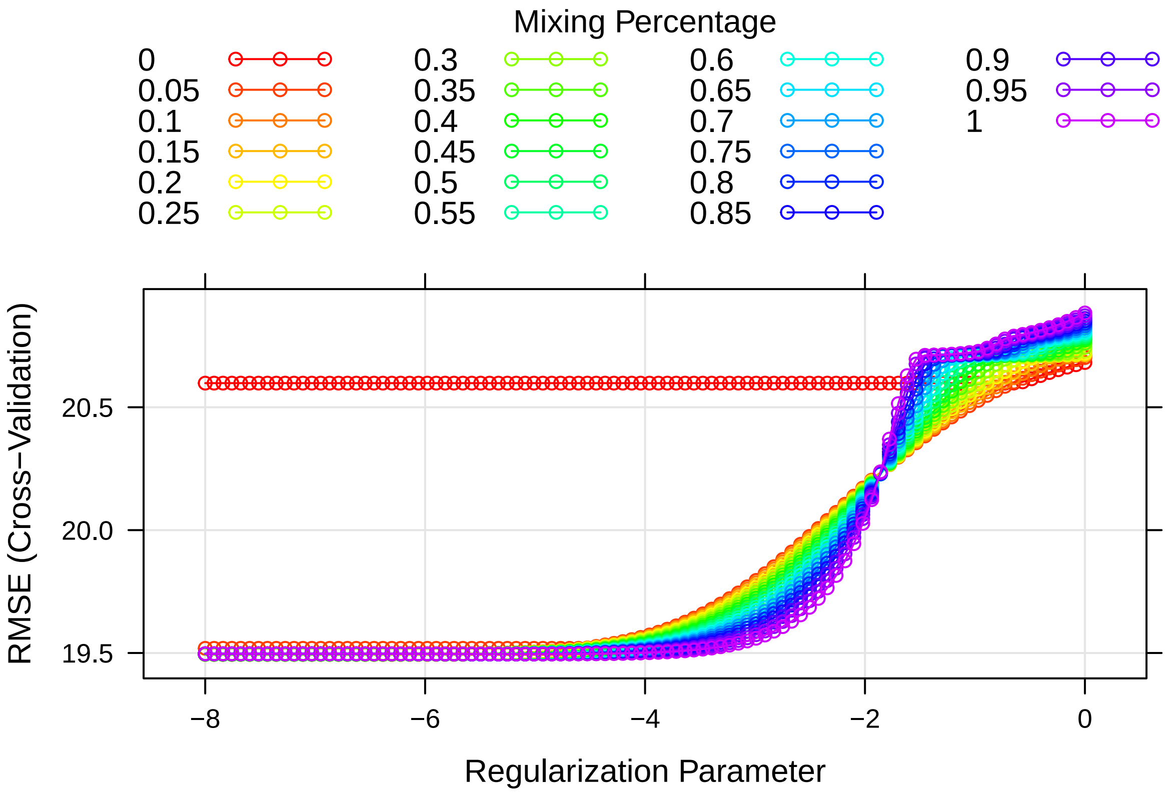


2.2 Elastic Net

Model assumptions:
(a) Combined Regularization: Elastic Net uses both L1 and L2 regularization, combining Lasso's variable

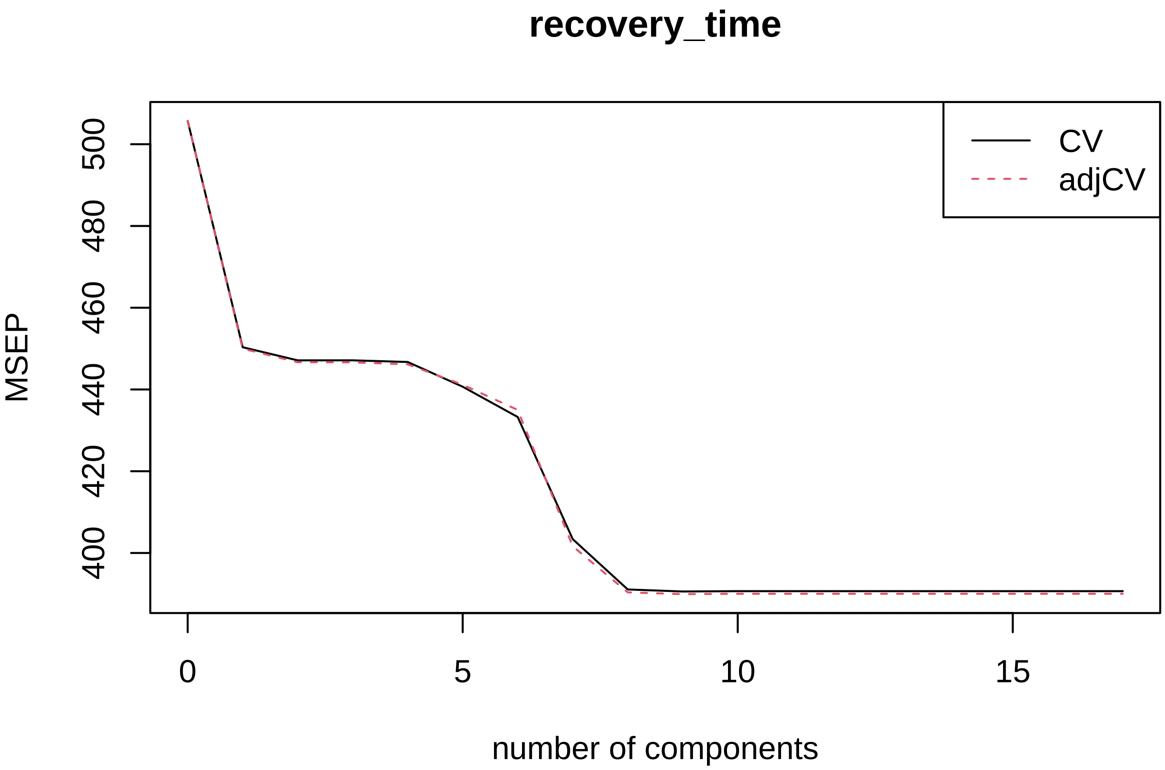
selection capability with Ridge regression's ability to handle highly correlated predictors.
(b) Adjusting Regularization Balance: Elastic Net has two regularization parameters, controlling the overall strength of regularization and the weight balance between L1 and L2 terms. This offers more flexible model tuning capability.





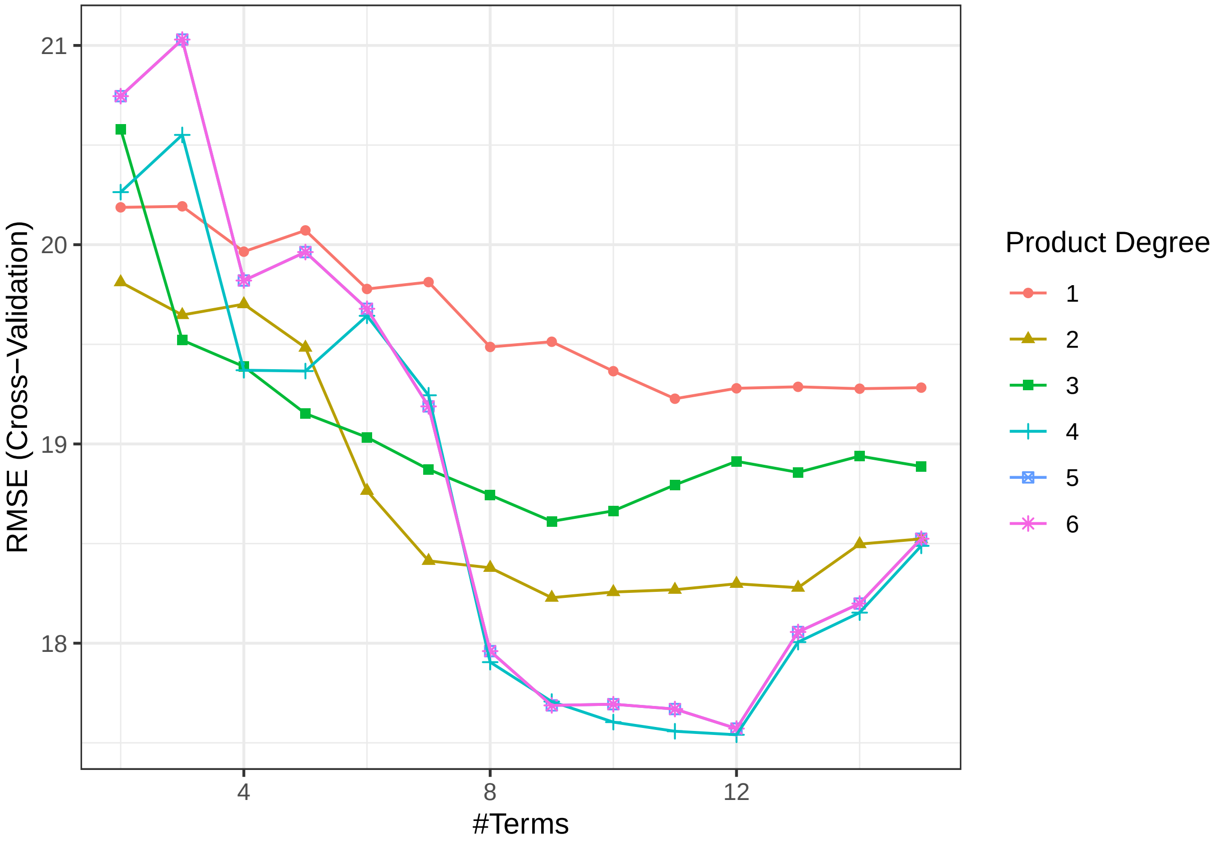
2.3 PLS

Model assumptions:
(a) Linear Relationship: PLS assumes a linear relationship between the independent variables and the response variable. It aims to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space.
(b) PLS assumes that the structure of the relationship between X and Y variables can be captured through a few latent structures. This is fundamental to reducing dimension and extracting the most relevant information from X that predicts Y.



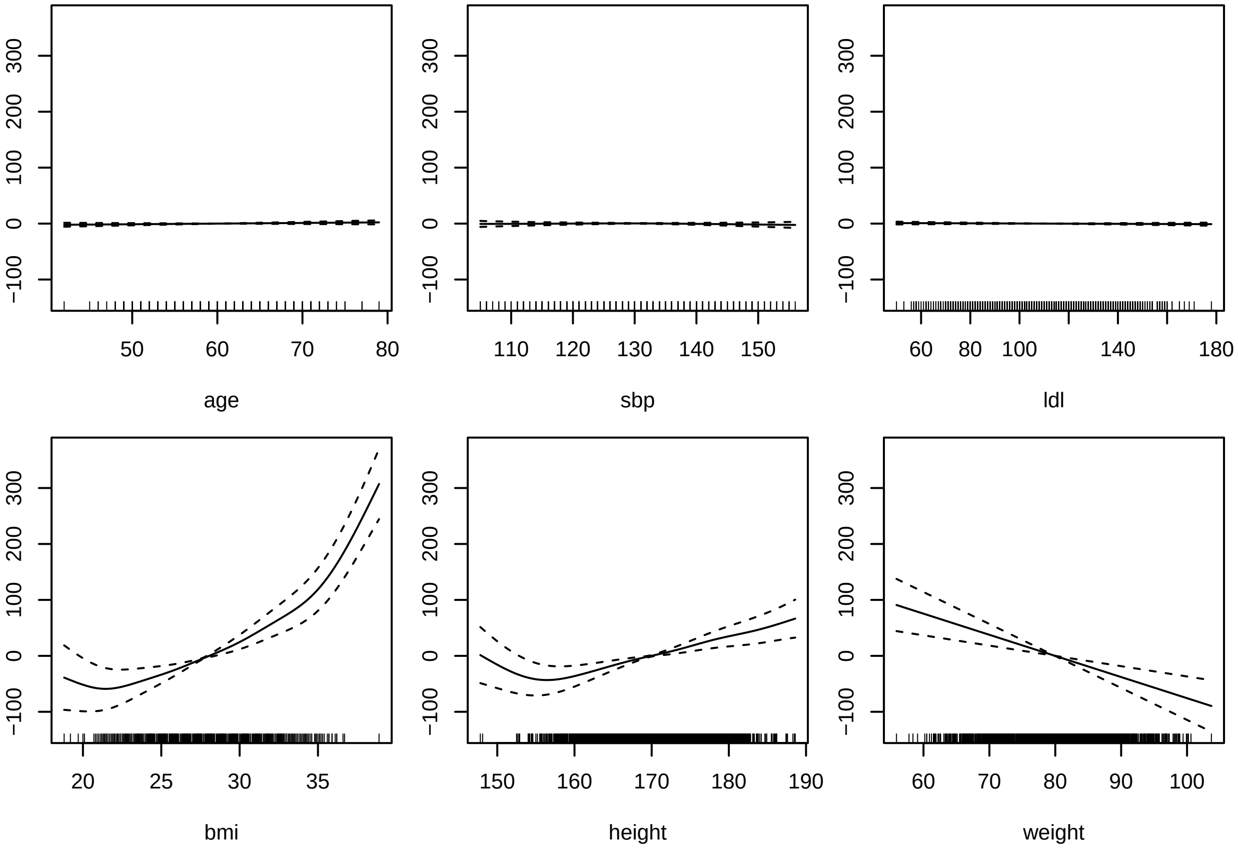
2.4 MARS

Model assumptions:
(a) Non-linearity and Interaction: MARS does not assume that relationships between the independent variables and the dependent variable are linear or follow a specific functional form. Instead, it adaptively fits piecewise linear regressions that can model complex non-linear relationships and interactions among variables.
(b) Distribution of Errors: MARS does not make specific assumptions about the distribution of error terms.



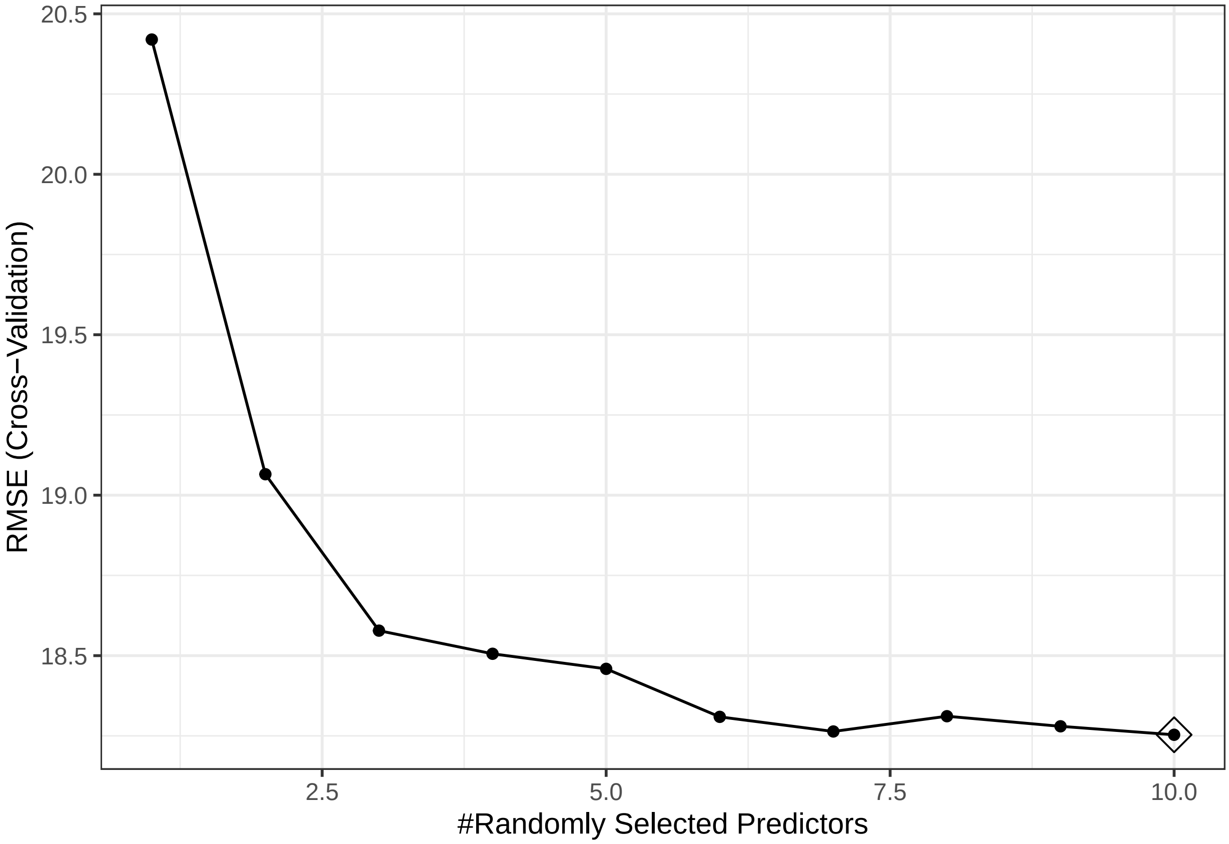
2.5 GAM

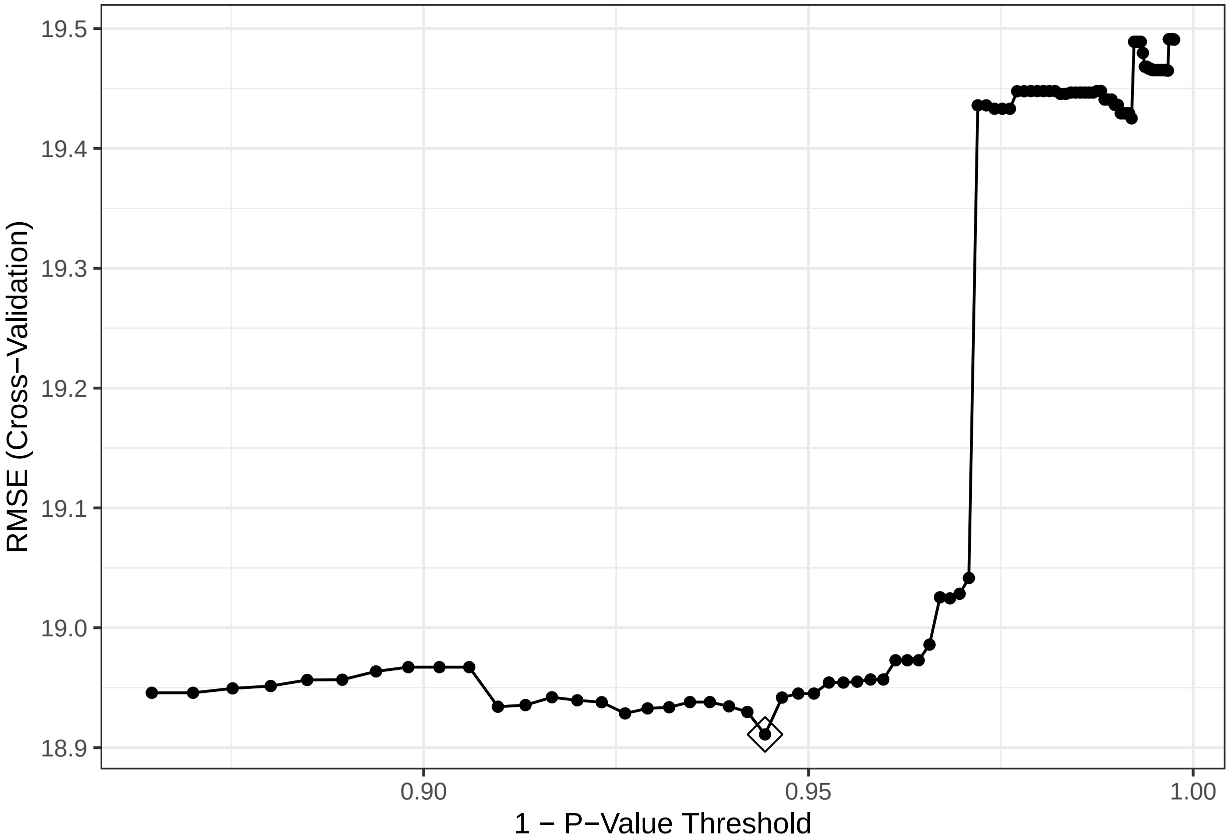
Model assumptions:
(a) Additivity: The effect of each predictor on the response is additive. The total effect on the response variable is the sum of the effects of each predictor, modeled by its own smooth function.
(b) Smoothness of the Predictor Functions: The relationships between the predictors and the response can be adequately modeled using smooth functions. The degree of smoothness is usually determined by the data and is controlled by smoothing parameters, which can be estimated from the data itself.



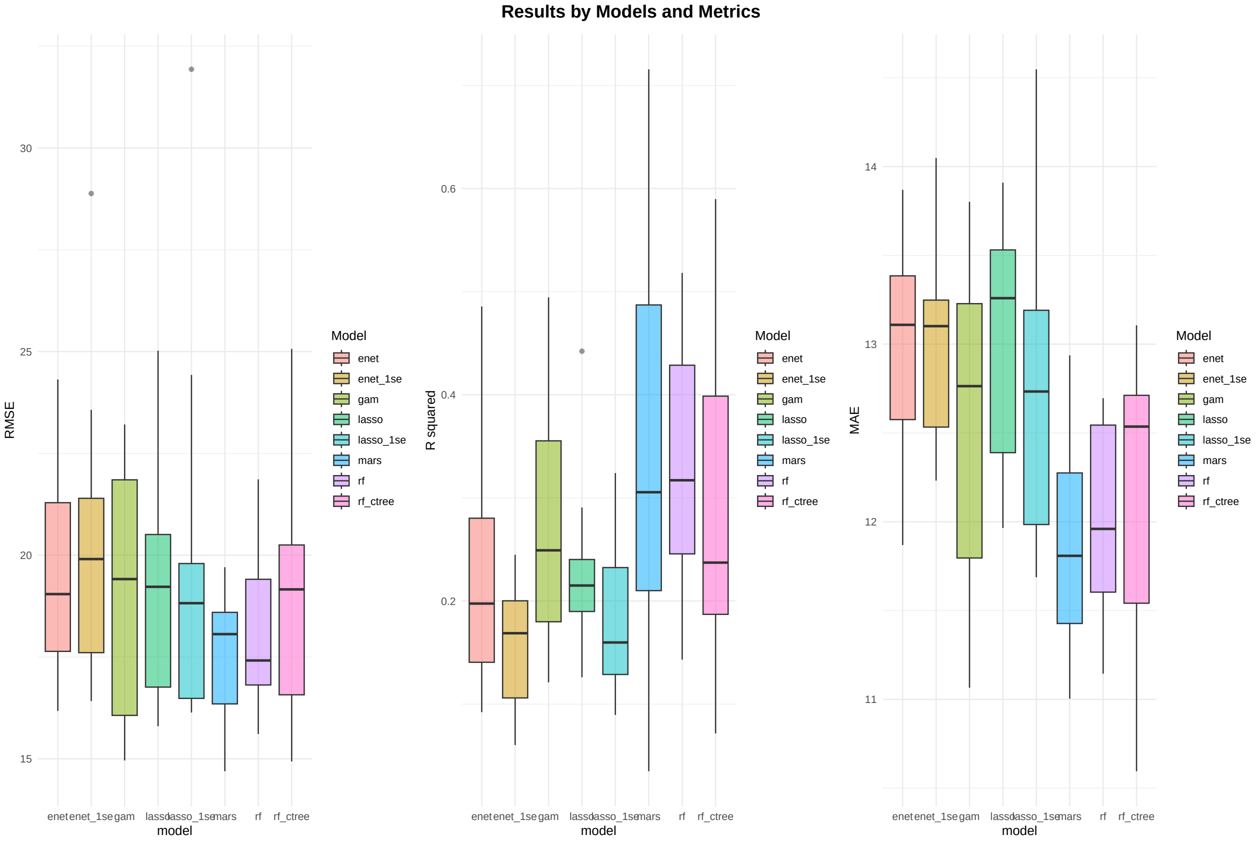
2.6 Random Forest

Model assumptions:


RMSE (Cross−Validation)
20.5
20.0
19.5
19.0
18.5
2.5
5.0
7.5
10.0
#Randomly Selected Predictors


RMSE (Cross−Validation)
19.5
19.4
19.3
19.2
19.1
19.0
18.9
0.90
0.95
1.00
1 − P−Value Threshold

2.7 Model Comparation



3 Result

4 Conclusion

5 Appendix