P8106 Data Science II Final Project
Identify risk factors for COVID-19 long recovery time by prediction model through statistical learning

Charles Chen mc5296
Hongru Tang ht2611
Jianting Shi js5095

May 10, 2023

1 Introduction and Background

To gain a better understanding of the factors that predict recovery time from COVID-19 illness, a study was designed to combine three existing cohort studies that have been tracking participants for several years. The study collects recovery information through questionnaires and medical records, and leverages existing data on personal characteristics prior to the pandemic. The ultimate goal is to develop a prediction model for recovery time and identify important risk factors for long recovery time.

2 Exploratory Data Analysis

There are 14 different variables available in the original dataset, which has 3601 records. To better understand and visualize the association between these variables and the outcome recovery time, these variables were first assigned proper data types. For example, categorical variables have been factorized, such as gender, race, hypertension, diabetes, etc. Scatter plot has been used to show the association of numeric variables with the outcome, while box plot is utilized to show the distribution of recovery time within different groups, such as gender and race, or levels of conditions such as the severity of COVID-19 infection or vaccination status at the time of infection. Correlation matrix and correlation plot are to explore the potential correlation and collinearity between variables, so as to provide extra information to determine whether to keep the variables in the final prediction model.
Correlation plot (Figure 1) shows that some of the variables are highly correlated with each other: positively correlated factors such as bmi and weight, SBP and age; negatively correlated factors such as bmi and height.

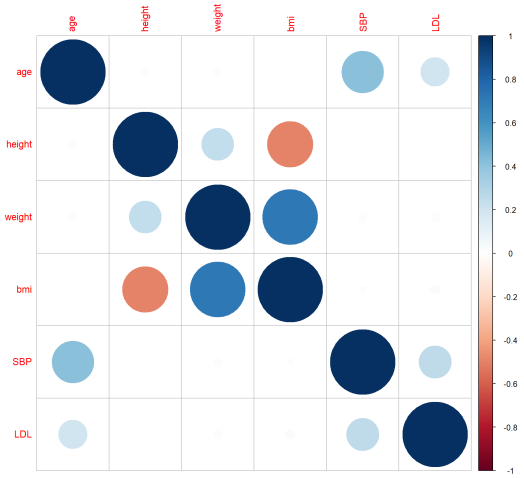

Figure 1: Correlation plot

Feature plot (Figure 2) shows that there are variables with high association with the outcome recovery time: numeric variables such as bmi/weight, height. Box plot (Figure 3) shows that there might be a significant difference of recovery time between/among the groups within some categorical variables, such as severity, vaccination, study B. Another set of box plots (Figure 4) to examine potential differences between groups of long or short recovery time after transforming the outcome into binary indicates the largest difference might lie in bmi index.

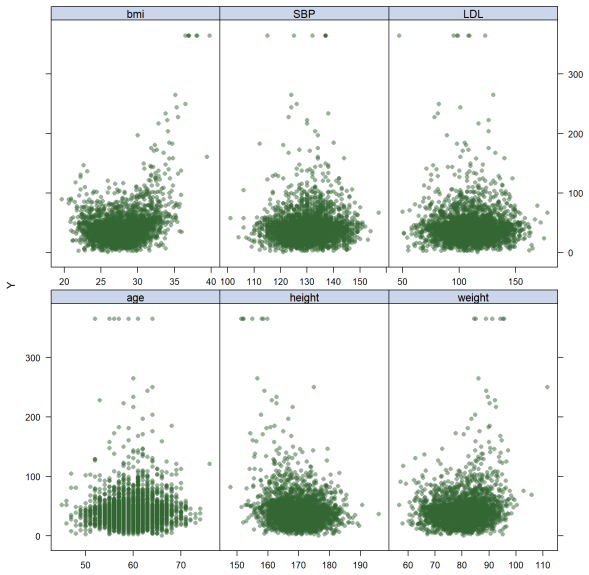

Figure 2: Relationship between recovery time and continuous variables

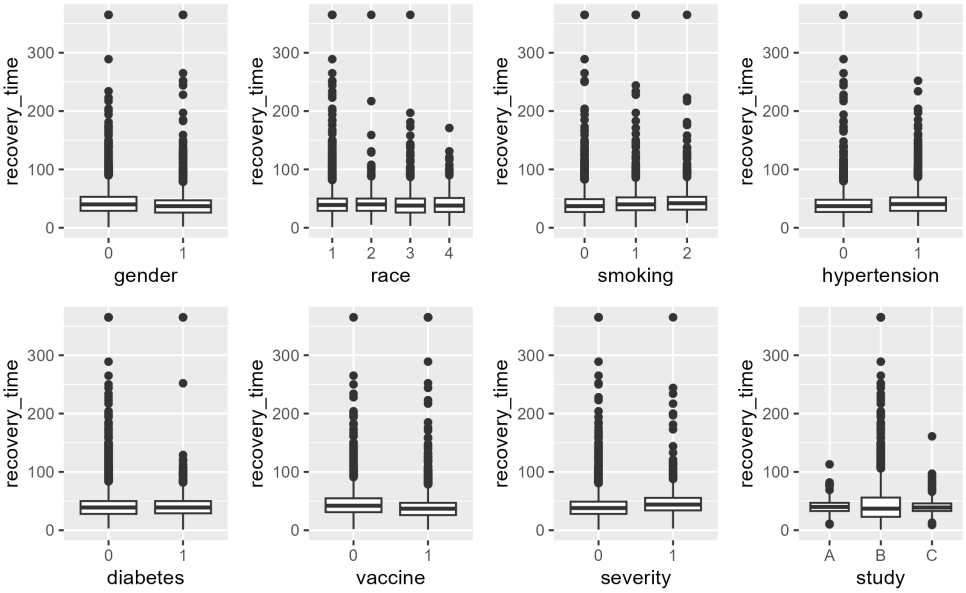

Figure 3: Relationship between recovery time and categorical variables

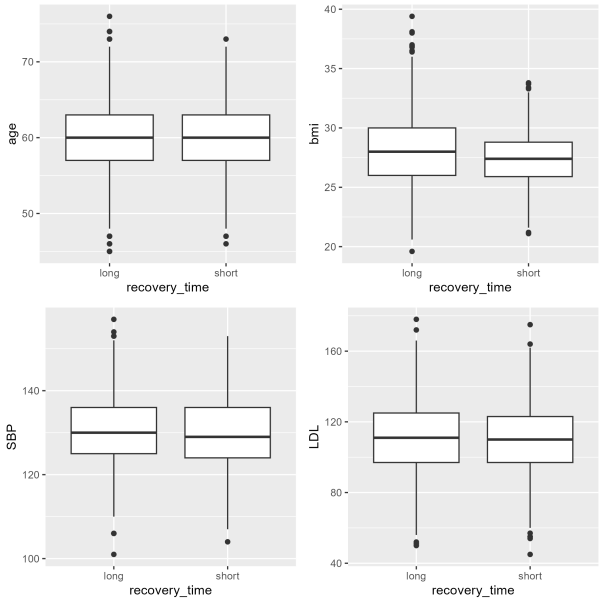

Figure 4: Relationship between binary recovery time and continuous variables

3 Model training on continuous recovery time

In order to find statistically significant relationships between variables and outcomes and make prediction, we trained following models on our dataset and made comparison.

3.1 Linear Regression

First, a linear regression is fitted assuming there is little or no multicollinearity in the data. The result shows the adjusted R-squared is about 0.2731. Only a portion of the variables is significant in the linear model, including gender, race3, smoking, height, weight, bmi, hypertension, LDL, vaccination, severity, and group B study. Since the variation inflation factor (VIF)is larger than 1, and based on the indications from the correlation plot, there is a potential correlation between some variables. So, the adjustment to the linear model needs to be made to increase the prediction performance.

3.2 Lasso, Ridge and Elastic-Net Regression

Ridge regression (Figure 7 in Appendix), lasso (Figure 8 in Appendix), and elastic net (Figure 9 in Appendix) methods are used to cope with groups of highly correlated predictors or perform feature selection. The tuning parameter lambda is 0.8387191 in the ridge model, alpha is 1 and lambda is 0.006737947 in the lasso model, alpha is 1 and lambda is 0.1353353 in the elastic net model. The ridge regression model has shrunken some of the non-significant coefficients in the linear model close to zero, while the significant ones from the linear regression model remain relatively large, such as bmi, weight/height, vaccination as well as study group B and severity. Lasso model has further shrinking the coefficients with more conservative p-values close to zero, such as age, race2, diabetes, SBP level, LDL level and study group C, indicating that these variables might not be good predictors. In elastic net model, the coefficient with the largest weight is consistent with the other two models, that is, bmi.

3.3 Partial Least Squares

Partial least squares (PLS) is also explored in the hope to identify dependent variables and thus to use this subset of latent variables to find the model with better prediction of the recovery time, especially via

PLS method. PLS method indicates there should be 11 components or transformed variables in the model (Figure 10 in Appendix).

3.4 Non-linear methods(GAM and MARS)

Non-linear methods such as Generalized Additive Model (GAM) and Multivariate Adaptive Regression Splines (MARS) are used to address the potential non-linearity of each variable. GAM identified 6 smoothing functions for age, SBP, LDL, bmi, height and weight, respectively. The partial effect plots (Figure 11 in Appendix) suggests there is obvious nonlinearity between BMI and the recovery time. BMI is the most weighted coefficients among all the others in GAM model. MARS model selected 17 of the 22 terms and 10 of the 18 predictors(Figure 12 in Appendix), most of which are interaction terms of bmi and other variables. Partial dependence plots(Figure 6) also shows a non-linearity between BMI and recovery.

3.5 Regression Tree

The main idea is to segment the predictor space into a number of simple regions, then fit a very simple model in each region. For our model, we consider a sequence of trees indexed by a non-negative tuning parameter α, it controls a trade-off between the subtree's complexity and its fit to the training data. By cross validation, we find that the best α is 0.01(Figure 13 in Appendix), and the corresponding tree size is 6(Figure 14 in Appendix).

3.6 Random Forest

Random forest provides an improvement over bagged trees by decorrelating the trees – reducing the variance when we average the trees. When building these decision trees, each time a split in a tree is considered, a random selection of m predictors is chosen as split candidates from the full set of p predictors. For regression: we train our method on the bth bootstrapped training set. Our goal is to average all the predictions to obtain:

$$f_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{*b}(x) \tag{1}$$

Random forest is decided by two parameters. The first is mtry which controls how many variables each tree used to predict. The second is minimal node size which controls the least number of predictors observed by every node point. We used cross-validation based on RMSE to conduct grid search. The mtry and minimal node size are 4 and 1 respectively(Figure 15 in Appendix).

3.7 Boosting

Boosting not only grows trees sequentially but also transforms each weak decision tree into a strong learner considering the error information from previous trees.
First, we fit a decision tree $f^b(x)$ with d split to the training data (X, r), and we update the tree by adding a shrunken version of the new tree:

$$f(x) = f(x) + \lambda * f^b(x) \tag{2}$$

Then we update the residuals:

$$r_i = r_i - \lambda * f^b(x_i) \tag{3}$$

In the end, we output the boosted model:

$$f(x) = \sum_{b=1}^{B} \lambda * f^b(x) \tag{4}$$

Boosting model is decided by 3 parameters. The number of trees B cannot be too large, otherwise the model will overfit. The shrink parameter λ balance the variance and bias. The number of split d in each tree Controls the complexity of the boosted ensemble. The best tuning parameter selected for Boosting are 2000, 0.005, 3 separately(Figure 16 in Appendix).

4 Model training on binary recovery Time

4.1 Logistic Regression

logistic Regression is the simplest model to predict binary outcomes. It has the form:

$$log(P(Y = 1|X)/1 - P(Y = 1|X)) = \beta_0 + \Sigma_i \beta_i x_i \tag{5}$$

We used other variables as predictors to fit a model to predict the probability that a subject belongs to a certain class. Here we used the default cut point p = 0.5. If the predicted probability of belonging to the long recovery time group is greater than 0.5, the subject will be predicted as a patient with long recovery time. We fit a logistic regression on the training set at first.

4.2 Linear Discriminant Analysis(LDA)

LDA is a statistical method that aims to find a linear combination of features that maximizes the separation between the classes, while also minimizing the within-class variance. The basic idea of LDA is to project the high-dimensional data onto a lower-dimensional space that preserves most of the class-related information. Given that the continuous variables in our dataset are normally distributed, we directly applied the LDA to our training set. The linear discriminant plot(Figure 17 in Appendix) shows that the long recovery time group is more spread out.

4.3 MARS

MARS is primarily a regression algorithm, but it can also be adapted for classification problems by using a method called binary recursive partitioning.For classification problem, MARS has the form:

$$log(P(Y = 1|X)/1 - P(Y = 1|X)) = \beta_0 + \Sigma_i \beta_i * h_i(x) \tag{6}$$

Where h(x) is a hinge function.
In binary recursive partitioning, the MARS algorithm constructs a piecewise constant function that estimates the probability of a binary outcome. The function is constructed by recursively partitioning the input space into smaller and smaller regions, each of which is associated with a constant value that estimates the probability of the outcome within that region. We applied the MARS model to our training dataset. We used cross-validation to choose the best tuning parameter. The best number of parameters is 11. and the degree is 1.(Figure 18 in Appendix) The cut point of the hinge function of BMI is 27.7 and 24. The hinge function of SBP has a cut point of 146.

4.4 Classification Tree

Classification tree is a kind of decision tree model. The tree is constructed recursively by selecting the best attribute or feature to split the data based on some criterion. The splitting continues until a stopping criterion is reached, such as the maximum depth of the tree, minimum number of samples in a leaf node, or a minimum improvement in the performance metric.
Classification tree model is easy to understand and interpret. It can handle mixed data types and missing data. So we also applied a classification model to our training dataset. The selected complexity parameter after cross validation is 0.0046(Figure 19 in Appendix). The final classification tree has a max depth of 12(Figure 20 in appendix).

5 Model Comparison and Result

5.1 Continuous recovery time

Comparison of the actual performance of the the models is achieved by obtaining the mean absolute error(MAE), residuals mean square error(RMSE) and R-squared from the cross-validation (Table 1). The

Boosting model have the lowest MAE, RMSE values and highest R-squared among all. Therefore, we select the boosting model as our final model.
The final model is also assessed through obtaining the MSE of test data. It is shown that boosting has a test error of 487.46.
In the boosting model, among these chosen variables, the importance of them in the model is ranked (Figure 5). The top 5 variables are bmi, study group B, height, vaccine and severity.

Model	MAE	RMSE	R-squared
Linear Model	16.773	25.104	0.263
Lasso	16.731	25.102	0.263
Ridge	16.871	26.686	0.167
Elastic Net	16.487	25.550	0.239
PLS	16.773	25.103	0.263
GAM	15.392	22.465	0.417
MARS	14.886	21.527	0.452
Regression Tree	15.063	22.000	0.428
Random Forest	14.752	21.339	0.453
Boosting	14.362	20.633	0.484

Table 1: Model comparison for continuous outcome based on MAE, RMSE and R-squared

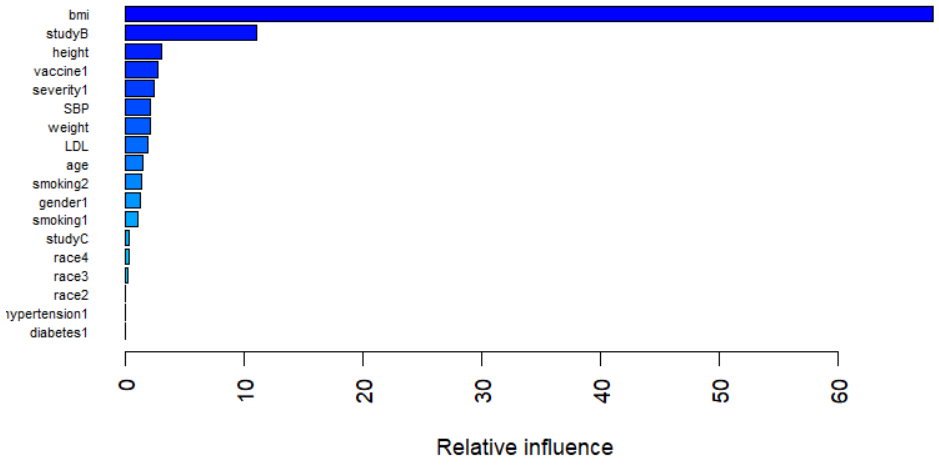

Figure 5: Features of importance in Boosting model

5.2 Binary recovery time

We used the ROC,sensitivity and specificity from the cross-validation to compare the models. The MARS model have the highest ROC and specificity values (Table 2) among all. Therefore we choose to use MARS as final model to predict the outcome.
The MARS model is also assessed through obtaining the Accuracy of test data. We also plotted ROC curve(Figure 21 Appendix) and generated a confusion Matrix (Table 3 in Appendix) on test data. It is shown that MARS model has an accuracy of 0.6968 and its AUC is 0.667, which is the highest among all

models. The final MARS model has the form:

$$\begin{aligned} log(P(short)/P(long)) &= 0.11 + 1.13 studyB - 0.48 * h(27.7 - bmi) + 0.57 vaccine - 0.70 severity \\ &+ 0.31 gender - 0.55 smoking2 - 0.36 smoking1 - 0.24 * h(SBP - 146) - 0.36 * h(bmi - 24) \\ &- 0.24 hypertension \end{aligned} \tag{7}$$

5.2.1 MARS model interpretation

The partial dependence plot (PDPs) (Figure 6) for BMI and the continuous recovery time from the boosting model, we could easily identify the non-linearity between these them. The recovery time does not significantly differ when the BMI index is in the range of 28.5 and 33.3. As the BMI index increases from 20 to 28.5, the recovery time decreases while it increases as the BMI index is beyond 33.3. The highest increasing rate is when the BMI index becomes larger than 35. While the recovery time does not differ when SBP is under 146 and becomes very high when it is larger than 146.

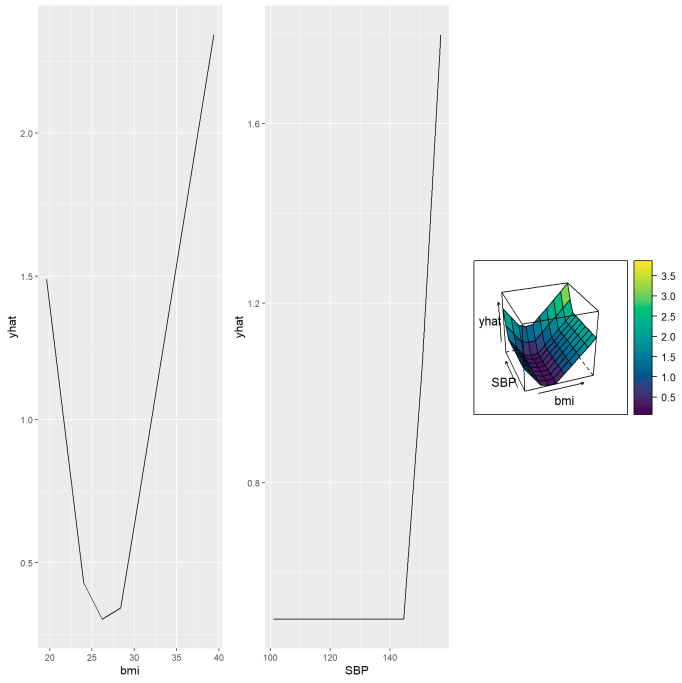

Figure 6: Partial Dependence Plot

In the MARS model, the association between the predictors and the outcome is implied by their coefficients. It suggests that the odds of shorter recovery time for those who are in the study group B is over 3 times the odds for those who are from other groups. Those who are vaccinated have the odds of recovering much faster over 1.7 times the odds of those who are vaccinated. The cut point at hinge function for bmi is at 24 and 27.7. It indicates that recovery time decreases as BMI index increases within range of 20-24 but increases as BMI index reaches beyond 27.7. Severity 1, smoking(1 and 2), hypertension are all with negative coefficients, suggesting that as severity gets higher or for those who have a history of smoking or are actively smoking, or those who with hypertension, they need more time to recover from COVID-19 infection.

6 Conclusion

It is suggested in the final models that the BMI index is the most informative and critical predictor for recovery time of COVID-19 infection. The association between recovery time as continuous variable and

Model	ROC	Sensitivity	Specificity
Logistic Regression	0.695	0.921	0.201
LDA	0.694	0.923	0.189
MARS	0.717	0.902	0.277
Classification Tree	0.661	0.897	0.260

Table 2: Model comparison for binary outcome based on MAE, RMSE and R-squared

BMI is non-linear but it's in general positively correlated with the increase of recovery time. Meanwhile, BMI seems to be the most significant predictors in other models with outcome .BMI also appears as an impactful predictor in groups for long or short recovery time. Thus, high BMI index could be a critical risk factors for COVID-19 long recovery time. Since weight and height are correlated with BMI index, they could also serve as predictors for COVID-19 time to recover. The other important predictors are vaccination status, the severity degree of the COVID-19 infection, the history of smoking, gender, and SBP level. With these factors combined, male who have not been vaccinated, with a history of smoking, with high BMI index and SBP level would have much higher risk of long recovery time of COVID-19 infection than others.In order to recover in a shorter time, we recommend actively vaccinating, exercising actively, and controlling diet to control weight. We also advocate for smoking cessation.
To be noticed, people who are from study group B tend to have longer recovery time but not from group A or C. It might be worth to take further investigation into the characteristics of this group of population to find out the actual factors which lead to the longer recovery time.

7 Appendix

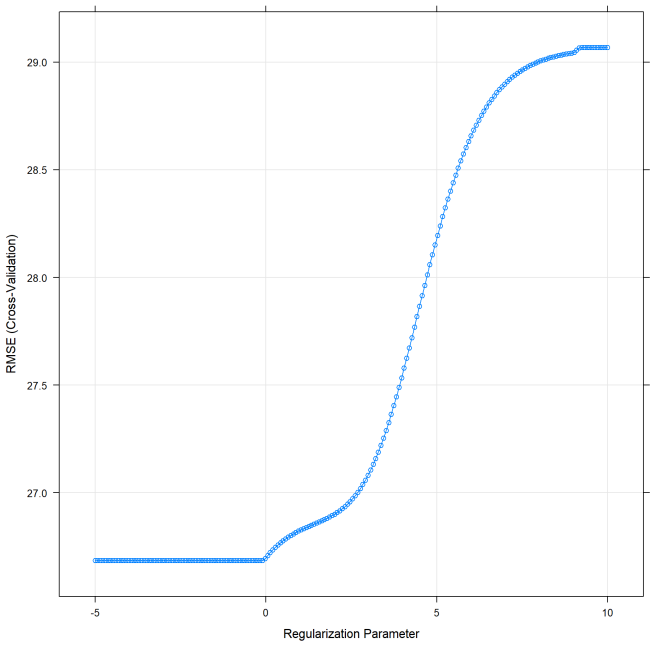

Figure 7: Model tuning for ridge regression

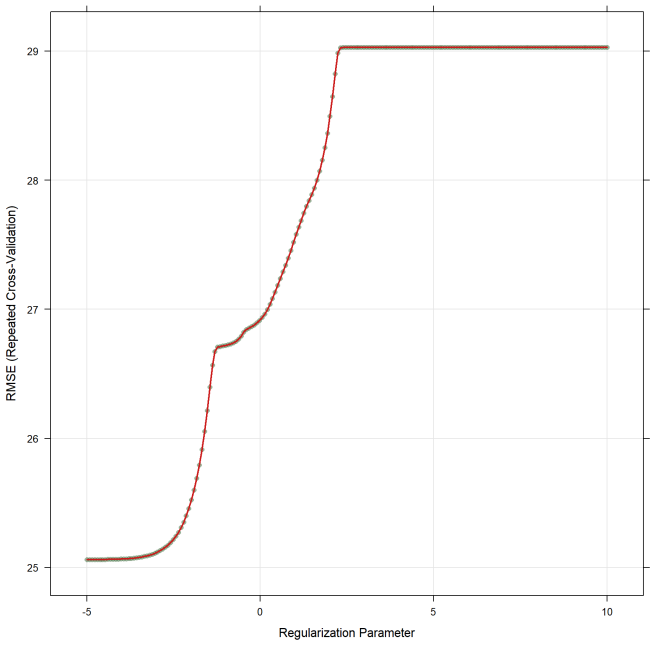

Figure 8: Model tuning for LASSO regression

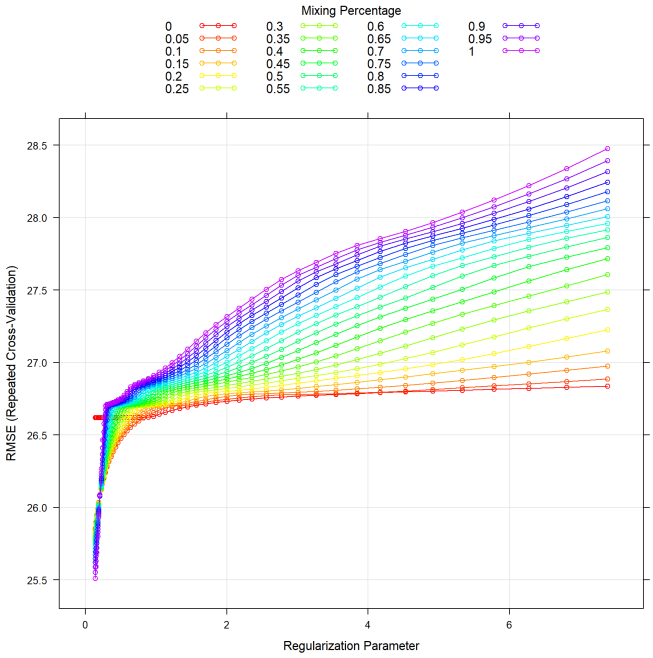

Figure 9: Model tuning for Elastic-Net regression

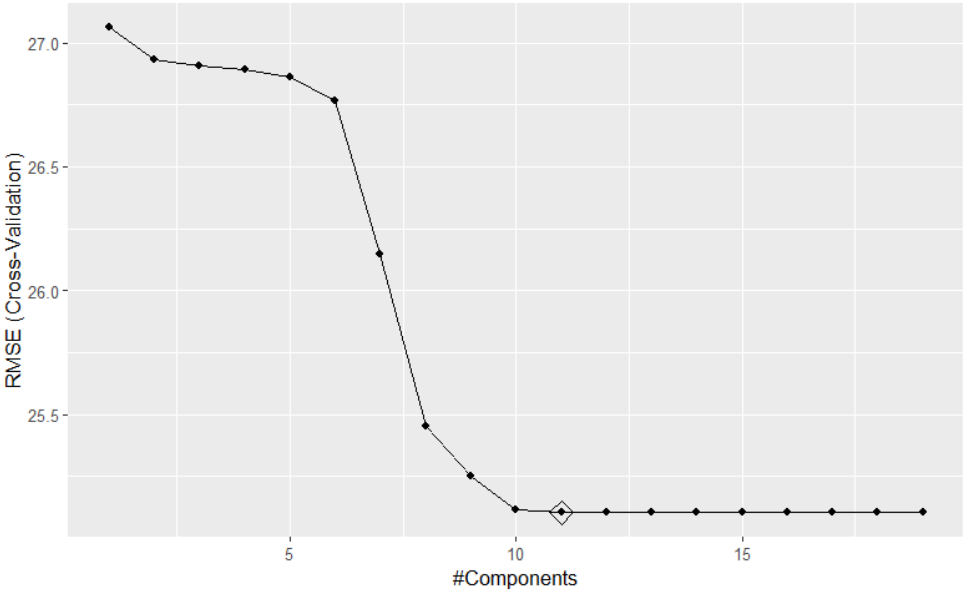

Figure 10: Model tuning for Partial Least Squares

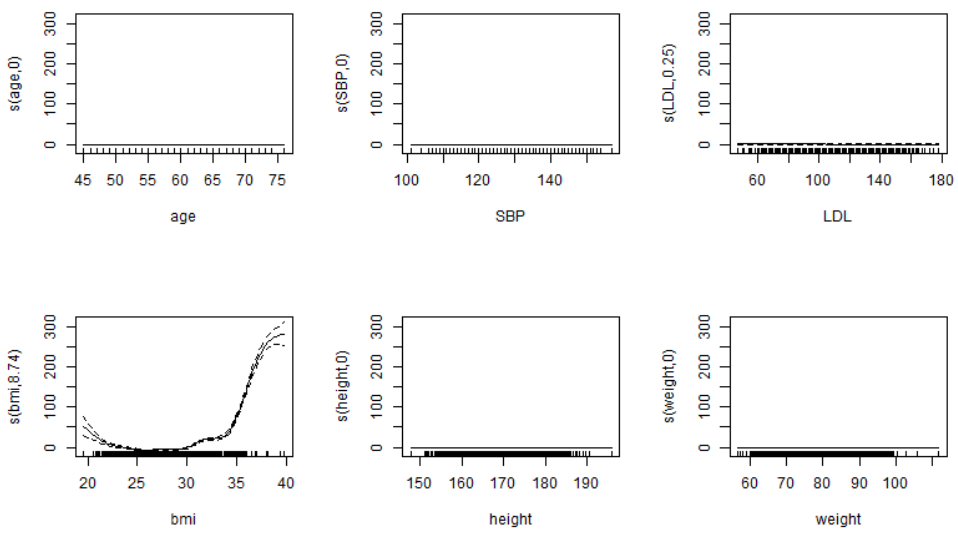

Figure 11: Smooth terms in GAM model

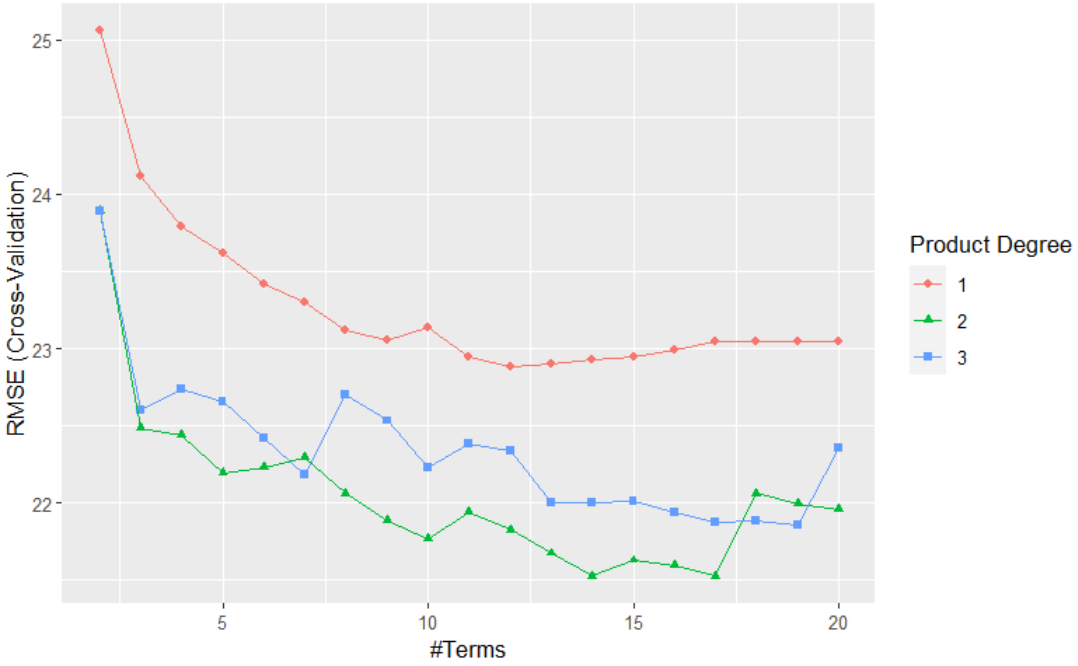

Figure 12: Model tuning for MARS

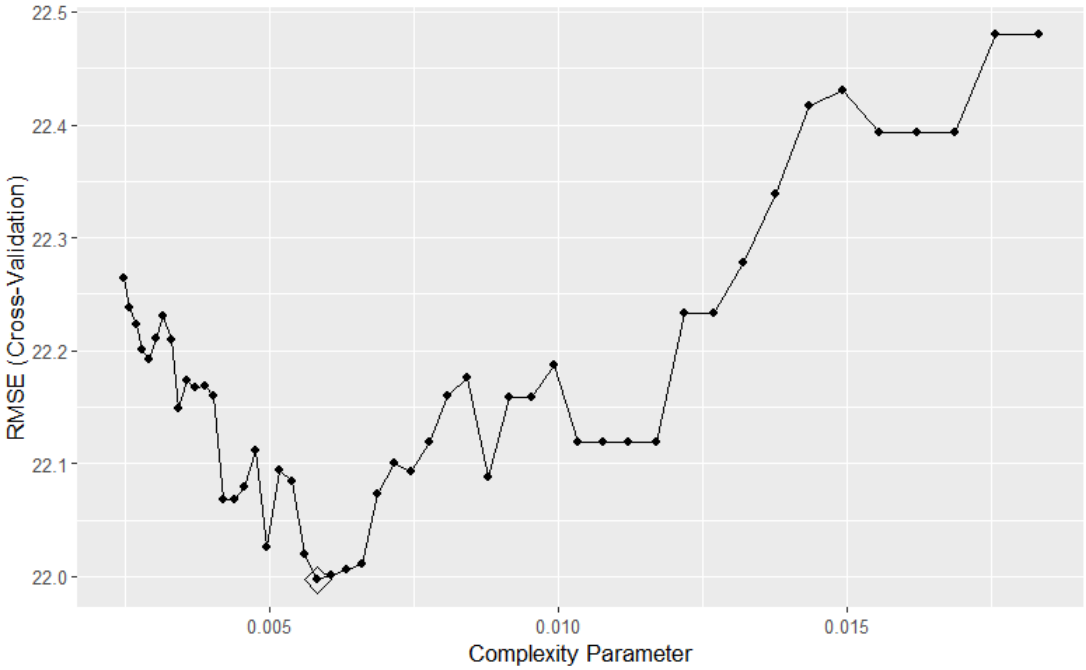

Figure 13: Model Tuning for Regression Tree

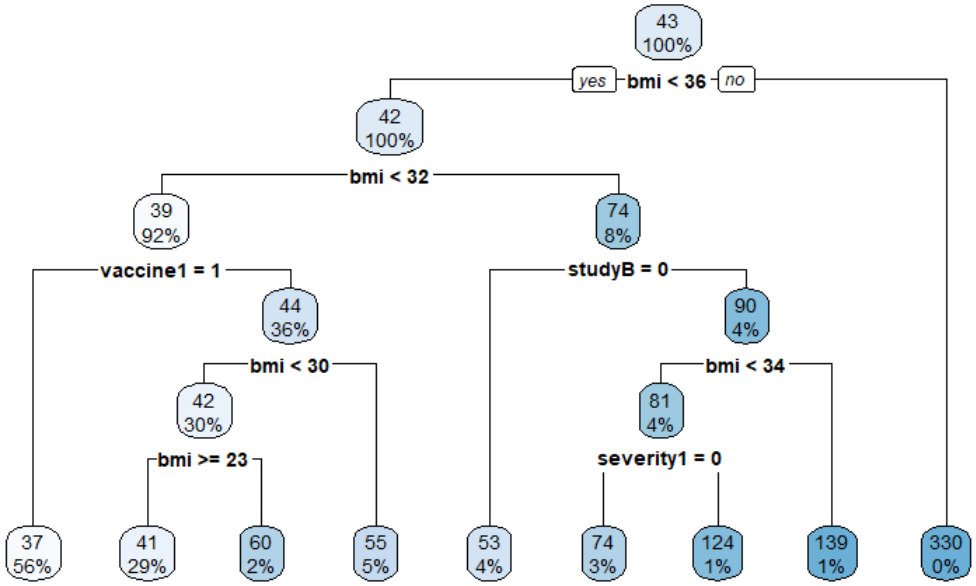

Figure 14: Regression Tree

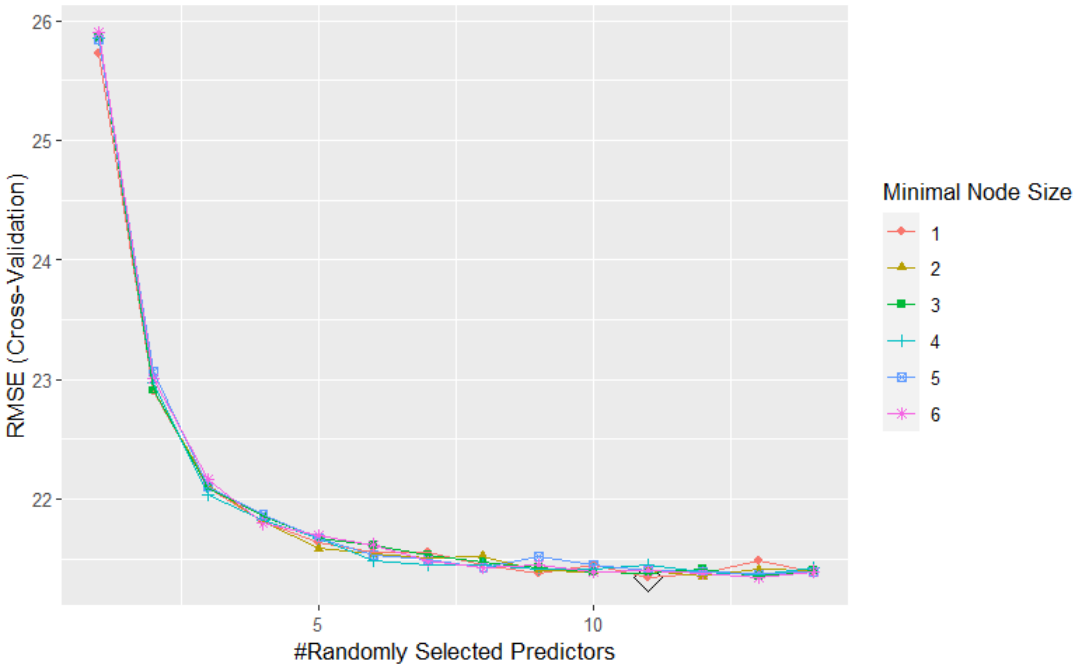

Figure 15: Model tuning for random forest

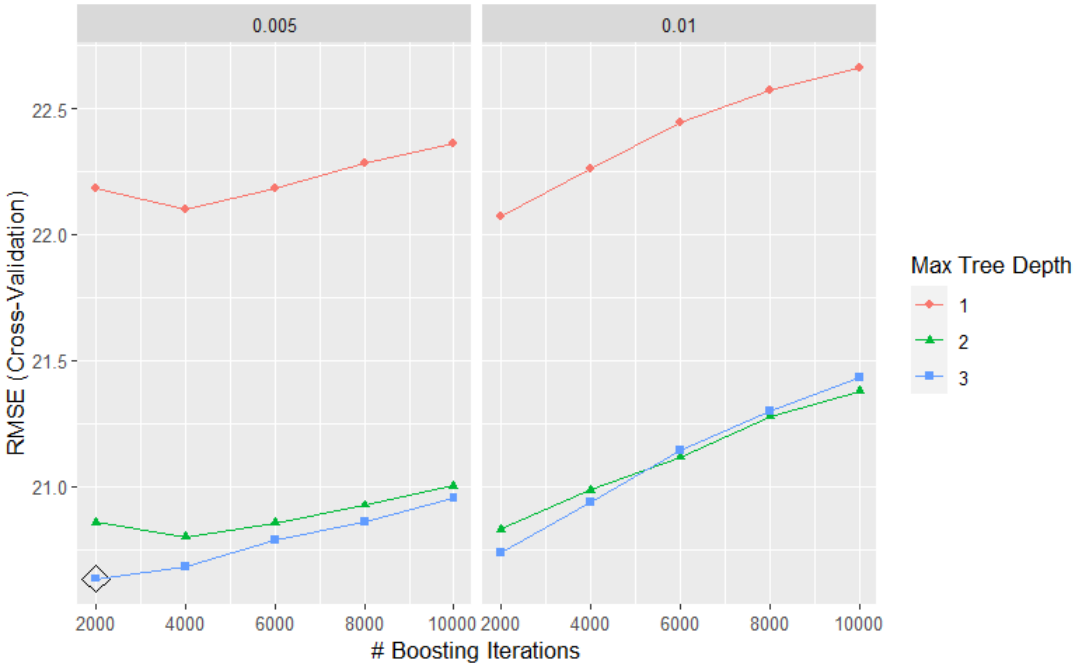

Figure 16: Model tuning for boosting

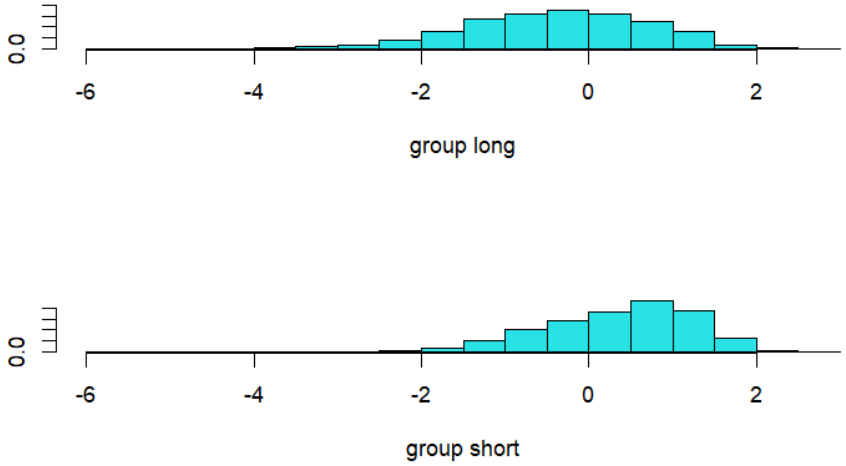

Figure 17: Linear discriminant plot

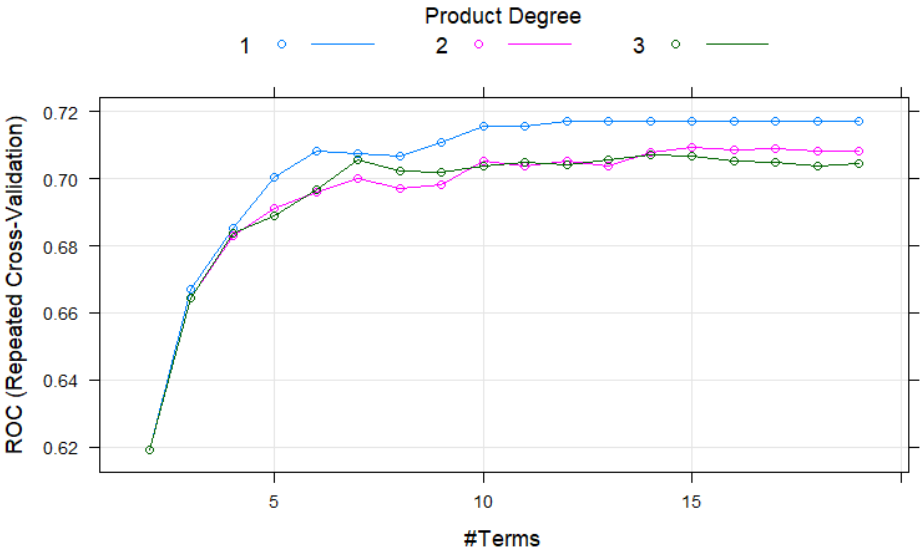

Figure 18: Model tuning for MARS

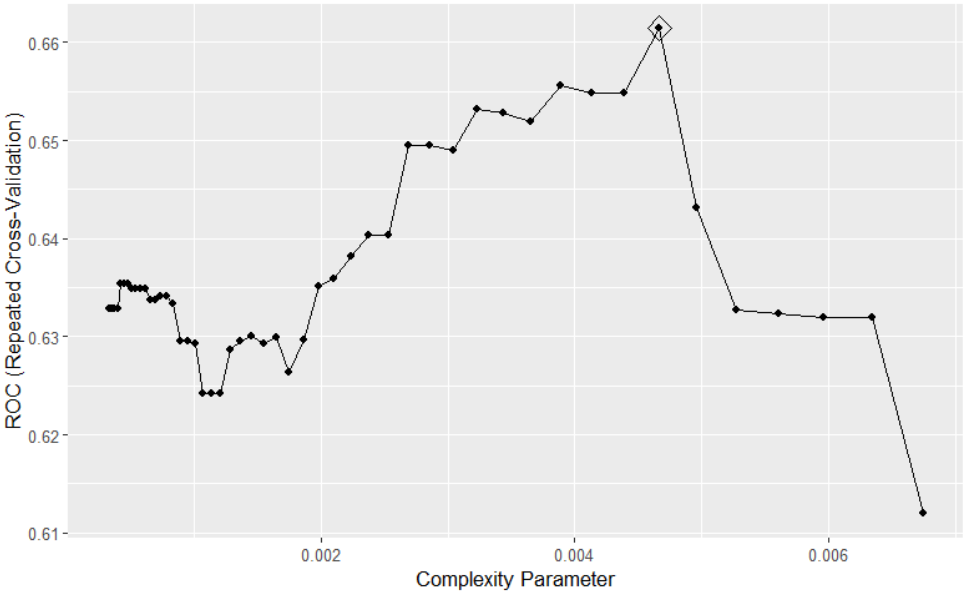

Figure 19: Model tuning for classification Tree

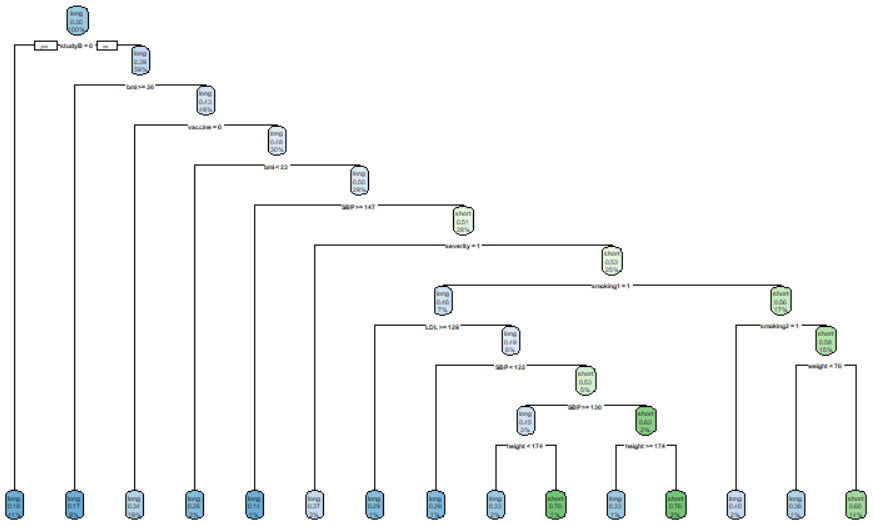

Figure 20: Classification Tree

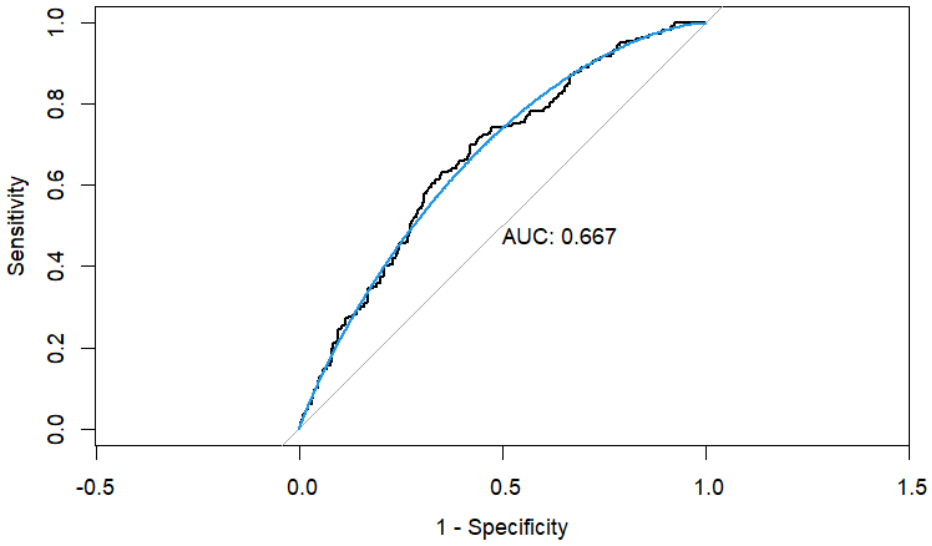

Figure 21: ROC curve for MARS

		Predicted Class	
		long	**short**
2***Actual Class**	**long**	444	160
	short	58	57

Table 3: Confusion Matrix