Prediction Models for Covid-19 Recovery Time and Risk factors for Long Recovery Time

Jingyi Yao (jy3269), Youlan Shen (ys3637), Zhengwei Song (zs2539)

2023-05-10

Contents

Introduction

The original dataset consists of the recovery time (measured in days) from COVID-19 and 15 other variables. Participant ID was excluded, assuming that the random effect among sample members is negligible. To indicate whether the recovery time is below or above 30 days, a new binary response variable was created. The study employs two types of models: the primary analysis aims to build regression models that predict the recovery time, while the secondary analysis focuses on classification models that predict whether the recovery time exceeds 30 days. Two midterm datasets in our group were merged to obtain a new dataset. The new dataset consists of 3569 observations, which was then split into 80% training data (with a sample size of 2879) and 20% testing data. Exploratory data analysis and model training will be conducted on the training dataset. The data partition was conducted using 2539 as the seed value, which will be consistently used in all subsequent analyses.

Exploratory analysis and data visualization

Continuous Predictors

The study includes six continuous predictors: age, height, weight, body mass index (BMI), systolic blood pressure (SBP), and low-density lipoprotein (LDL) cholesterol. Figure 1-a and 1-d demonstrate that the continuous predictors have a relatively symmetrical and approximately normal distribution, irrespective of whether the response variable is binary or continuous. Scatter plots in Figure 1-b show a nonlinear relationship between BMI and recovery time. The slopes of the fitted lines between SBP, LDL, height, weight, age and recovery time are close to zero, indicating a weak linear relationship between them.

To prepare for model training, it is necessary to analyze the correlations between predictors. Figure 1-c shows a strong positive correlation between BMI and weight, and a negative correlation between BMI and height. SBP vs. age, LDL vs. age, and SBP vs. LDL are moderately correlated. Highly correlated predictors can lead to issues such as multicollinearity, which may affect the prediction performance of the model. Therefore, further model training with cross-validation (CV) is required to address this issue.

Categorical Predictors

The study includes eight categorical predictors: gender, race, smoking status, hypertension status, diabetes status, vaccine status, severity of COVID-19 infection, and the study group (A/B/C) to which the participant belongs.

Figure 2 presents the distribution of continuous recovery time by each categorical predictor using box plots. There are no noticeable differences in the distribution of recovery time across the subclasses of each predictor. Thus, further model training is required to investigate potential associations between the categorical predictors and recovery time.

Figure 3-a displays a bar chart that shows 2005 people in the training dataset took more than 30 days to recover from COVID-19, while 874 people recovered within or less than 30 days. Moreover, Figure 3-b demonstrates that the binary outcome has the same distribution pattern among all predictors, which means that in each category, there are more people with a recovery time of over 30 days. The distribution appears to be slightly imbalanced, which may have some impact on the classification results.

Model Training

Regression

In order to determine the best model for predicting Covid-19 recovery time as a continuous outcome, we applied Linear Model (LM) and trained four other models, including Elastic Net, Multivariate Adaptive

Regression Splines (MARS), Regression Tree, and Random Forest. All model training processes were based on 10-fold cross validation to obtain the minimum CV error (RMSE, Root Mean Square Error). After cross validation, we selected the optimal tuning parameters for each model. Test errors (test RMSE obtained by applying the final model to the test data) were also reported for each model. We set the seed as `2539`.

1. Linear Model (LM)

LM assumes the relationship between the predictors and the outcome is linear. The observations are independent of each other and the residuals have constant variance across all values of the predictors. There is no tuning parameter for LM. Applying the LM model to test data, we obtained a test error of 22.57.

2. Elastic Net

The Elastic Net model also assumes a linear relationship between predictors and the response variable. It introduces both L1 penalty term from LASSO and L2 penalty term from Ridge Regression to the objective function. Compared to LASSO, Elastic Net is more robust in handling highly correlated predictors.

During the cross-validation process, we explored different values for tuning parameters α and β. We tried 21 candidate values for α (ranging from 0 to 1, with 0 indicating a Ridge model and 1 indicating a LASSO model) and 100 candidate values for β (ranging from e^{-8} to e^{-2}). The optimized parameters that minimized the cross-validation error were found to be $\alpha_1 = 0.15$ and $\beta_1 = 0.0055$. All predictors were retained in the final model, which achieved a test error of 22.48.

3. Multivariate Adaptive Regression Splines (MARS)

MARS assumes that the relationship between predictors and the response is piecewise linear, with the algorithm automatically selecting cut points. We conducted cross-validation and optimized the degree of features (from candidates 1, 2, 3) and the number of terms (from candidates 2, 3, ..., 25) to obtain the model with the minimum CV error. The optimized model identified a product degree of 1 and the number of terms as 16. The test error for this model is 21.07.

4. Regression Tree

Regression Tree is a recursive binary splitting method used to construct a large tree from the training data. It partitions the predictor space into a number of simple regions and fits a basic model in each region. Regression Tree assumes a non-linear relationship between the predictors and the outcome.

After cross-validation, we identified the optimal tuning parameter for the model as a complexity parameter of 0.0028 (from 50 candidates, split evenly from e^{-8} to e^{-4}). The test error was found to be 20.91.

5. Random Forest

Random Forest builds multiple decision trees on bootstrapped training samples. However, unlike traditional decision trees, at each split in a tree, a random subset of predictors is selected as split candidates from the full set of predictors. By using Random Forest, we can significantly improve prediction accuracy compared to a single tree. This algorithm assumes independence between observations, a non-linear relationship between outcome and predictors, roughly equal numbers of examples for each class, and normalized or standardized predictors.

We used variance as the split rule and optimized the model using cross validation. The optimal tuning parameters were 5 predictors from all predictors and a minimal node size of 6. The test error for this model is 21.04.

Classification

To identify the optimal model for predicting Covid-19 recovery time as a binary outcome, we evaluated six models through model training: logistic regression, MARS, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Random Forest, and AdaBoost. All models were trained using 10-fold cross validation to maximize the training accuracy. Following cross validation, we selected the optimal tuning parameters for each model. Additionally, we reported the test accuracy obtained by applying the final model to the test data for each model. The seed was set to `2539`.

1. Logistic Regression

Logistic regression shares the same assumptions with LM, except the first one, where Logistic regression assumes that the relationship between the predictor variables and the log odds of the response variable is linear. There is no tuning parameter for Logistic Regression. The model has a test accuracy of 69.78%.

2. MARS

In classification, the goal of MARS is to predict the class label of each observation based on a set of predictor variables. One of the key assumptions of MARS in classification is that the classes are separable by piecewise linear functions. Additionally, MARS assumes that the class labels are mutually exclusive and that there is no overlap between classes. This means that MARS is not suitable for multi-label classification problems where an observation can belong to more than one class.

After conducting cross-validation and obtaining the maximum CV accuracy, we optimized two tuning parameters: the degree of features (from candidates 1, 2, 3) and the number of terms (from candidates 2, 3, ..., 25). The optimized model identified a product degree of 2 and the number of terms of 18. The test accuracy of this model is 69.36%.

3. Linear Discriminant Analysis (LDA)

LDA assumes that the predictors are normally distributed and that they have equal covariance for all classes. Unlike Logistic Regression, LDA can be quite robust to deviations from normality. The model achieved a test accuracy of 69.64%.

4. SVM-Radial Kernel

Support Vector Machine-Radial Kernel is an efficient algorithm for learning nonlinear functions and assumes that the data should be separable, allowing us to draw nonlinear boundaries between two classes. It also assumes independence between observations.

After conducting cross-validation, we obtained the optimal tuning parameters of cost = 1.0851 (from 50 candidates, split evenly from e^{-2} to e^{4}) and sigma = 0.0363 (from 20 candidates, split evenly from e^{-7} to e^{-2}), which were tuned over both cost and sigma, resulting in a test accuracy of 71.17%. In a second round of tuning, we optimized only the cost parameter while using a single value of sigma based on kernlab's sigest function, resulting in the optimal tuning parameter of cost = 1.1710 (from 20 candidates, split evenly from e^{-3} to e^{3}) and a test accuracy of 71.03%.

5. Random Forest

We also utilized Random Forest for classification. After conducting cross-validation, we determined that the optimal tuning parameters were 3 predictors from all 14 predictors and a minimal node size of 18 (from candidates 6 to 20 by 2) under the Gini index. The resulting test accuracy was 71.73%.

6. Adaboost

AdaBoost is an algorithm that fits classification trees to weighted versions of the training data and updates the weights to better classify previously misclassified observations. It assumes independence between observations, a non-linear relationship between the outcome and predictors, roughly equal numbers of examples for each class, and that the predictors are well-normalized or standardized.

After cross-validation, we determined the optimal tuning parameters to be: the number of trees, B, set to 4000 (from 2000, 3000, 4000, 5000); the shrinkage parameter, λ, set to 0.001 (from 0.001, 0.002, 0.003); and the interaction depth, d, set to 4 (from 1, 2, ..., 10). Using these parameters, the test accuracy was found to be 72.14%.

Results

Regression

To compare multiple models, we use the 10-fold cross-validation resampling error (RMSE) as the criterion. In Figure 4, the five models are listed from top to bottom as Regression Tree, Random Forest, MARS, Linear Model and Elastic Net. As shown, Random Forest achieves the lowest mean cross-validation error

at approximately 22.41. Therefore, we select the Random Forest model for the primary analysis, treating recovery time as a continuous outcome. Applying the trained model to the test dataset, we obtained a test error (test RMSE) of 21.04.

To better understand the Random Forest model, we generated a variable importance plot along with a partial dependence plot, which are shown below. Figure 5-a reveals that BMI, weight, and height are the top three factors influencing COVID-19 recovery time. Figure 5-b shows that, controlling for the effects of other predictors, there is a decrease in recovery time as BMI increased from around 20 to around 25 kg/m^2. However, recovery time then increases sharply as BMI exceeds 30. The recovery time exhibits a similar decreasing-then-increasing trend in both weight and height. It decreases as weight increases from 60 to 80 kg or as height increases from 150 to 175 cm. Then, it increases as weight increases from 80 to 100 kg or as height increases from 175 to 190 cm. In terms of SBP, the recovery time increases as SBP increases from 125 to 132 mm/Hg, but decreases as SBP increases from 132 to 145 mm/Hg. The joint partial dependence plot of SBP and BMI shows that, for any given value of SBP, the recovery time increases with an increase in BMI. Furthermore, participants in study group B tend to have a longer recovery time compared to the other two study groups. Additionally, in the model, the estimated recovery time jumps from 42 to 50 days as infection severity occurs.

Classification

In the secondary analysis, we compared multiple models using the 10-fold CV resampling accuracy as the criterion. Figure 6 presents the seven models tested, including Logistic Regression, MARS, LDA, SVM with two versions of the Radial kernel, Random Forest, and AdaBoost. The results indicate that AdaBoost achieved the highest mean cross-validation accuracy, around 0.7207. Consequently, we selected the AdaBoost model for the secondary analysis, treating recovery time as a binary outcome. Applying the trained model to the test dataset, we obtained a test accuracy of 0.7214.

To gain a better understanding of the AdaBoost Model, we created a variable importance plot and a partial dependence plot, which are presented below. Figure 7-a shows that BMI and study group B are the two most important predictors for COVID-19 recovery time. In Figure 7-b, we can observe the probability of experiencing a longer recovery time (over 30 days) for each variable while keeping all other variables constant. The probability of longer recovery time decreases as BMI increases from around 22 to 25 kg/m^2, but it sharply increases once BMI exceeds 30. As for LDL, the probability sharply increases starting at around 55 mg/dL and then remains stable. However, since there is a lack of data for LDL between around 55 to 90, a more precise trend within this interval is not available. In terms of SBP, the probability of long recovery time generally increases before SBP reaches 135 mm/Hg, and slightly decreases afterwards. The joint partial dependence plot of SBP and BMI shows that, for any given value of SBP, the probability of longer recovery time decreases as BMI increases from around 20 to 30, and then it increases as BMI exceeds 30. In contrast to the Random Forest model, the participants in study group B exhibits a lower probability of experiencing a longer recovery time compared to the other two study groups. Additionally, in the model, the estimated probability of long recovery time jumps from 0.69 to 0.82 as infection severity occurs.

Conclusions

The Random Forest model was selected as the prediction model for continuous recovery time due to its lowest resampling mean error, while Adaboost was chosen as the prediction model for the binary response variable recovery time because of its highest resampling mean accuracy. Comparing the two models, we may conclude that models that do not assume a linear relationship between outcome and predictors generally provide better prediction results. Based on the model results, important risk factors for longer COVID-19 recovery time can be identified.

Risk factors

1. BMI

The participant's BMI is a significant risk factor for longer recovery time when $BMI > 30\ kg/m^2$ or when $BMI < 24$. Therefore, individuals who are in the obese range*[1] ($BMI > 30\ kg/m^2$) or in the underweight range*[1] ($BMI > 18.5\ kg/m^2$) are at higher risk of experiencing a prolonged COVID-19 recovery time, adjusting for the other predictors. Additionally, weight and height are also important factors that contribute to recovery time, however, adult BMI*[3] contains the information of both weight and height. As a result, the BMI is identified as a risk factor, indicating that individuals who are either underweight or overweight are more likely to take longer to recover from COVID-19.

2. SBP

People with Stage 2 hypertension*[2] ($SBP > 140\ mm/Hg$), Stage 1 hypertension*[2] ($SBP > 130\ mm/Hg$) or elevated SBP*[2] ($SBP > 120\ mm/Hg$) are at risk of long COVID-19 recovery time, holding other predictors fixed.

3. Infection Severity

Severe infection gained relatively high scores in the two final models' variable importance plots. Thus, being severely infected by COVID-19 can be considered as a risk factor for longer recovery time.

4. Study Group B

Although the effect of study group B on recovery time is opposite in the two final models, study group B gained high scores in both models' variable importance plots. Therefore, study group B may also be an important factor for recovery time compared to the other two study groups. However, as there is a lack of information on how the study groups were partitioned, we may suspect that individuals in group B share certain common features that impact COVID-19 recovery time. Further research is needed to better understand the characteristics of individuals in group B that may contribute to their longer recovery time.

Predictions

Our analysis suggests that individuals who fall under the obese or underweight BMI ranges, have high SBP ($SBP > 120\ mm/Hg$), or suffer from severe COVID-19 infections are more likely to experience a longer recovery time. Given these findings, it is crucial to allocate more policy support and healthcare resources towards these populations for the future prognosis of COVID-19.

Reference

[1] Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults—The Evidence Report. National Institutes of Health.Obes Res. 1998; 6 (Suppl 2):51S–209S.

[2] Whelton PK, Carey RM, Aronow WS, Casey DE Jr, Collins KJ, Dennison Himmelfarb C, DePalma SM, Gidding S, Jamerson KA, Jones DW, MacLaughlin EJ, Muntner P, Ovbiagele B, Smith SC Jr, Spencer CC, Stafford RS, Taler SJ, Thomas RJ, Williams KA Sr, Williamson JD, Wright JT Jr. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol. 2018 May 15;71(19):e127-e248. doi: 10.1016/j.jacc.2017.11.006. Epub 2017 Nov 13. Erratum in: J Am Coll Cardiol. 2018 May 15;71(19):2275-2279. PMID: 29146535.

[3] Garrow, J.S. & Webster, J., 1985. Quetelet's index (W/H2) as a measure of fatness. Int. J. Obes., 9(2), pp.147–153.

Figures

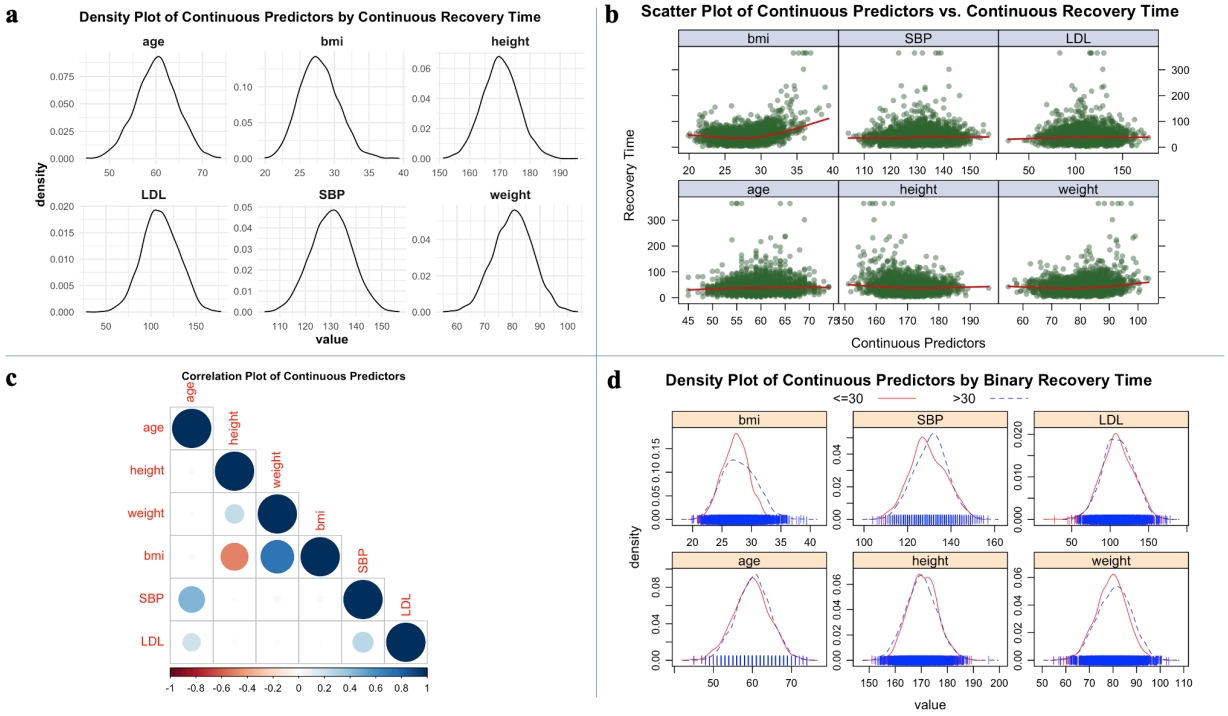

Figure 1: Visualization of Continuous Predictors

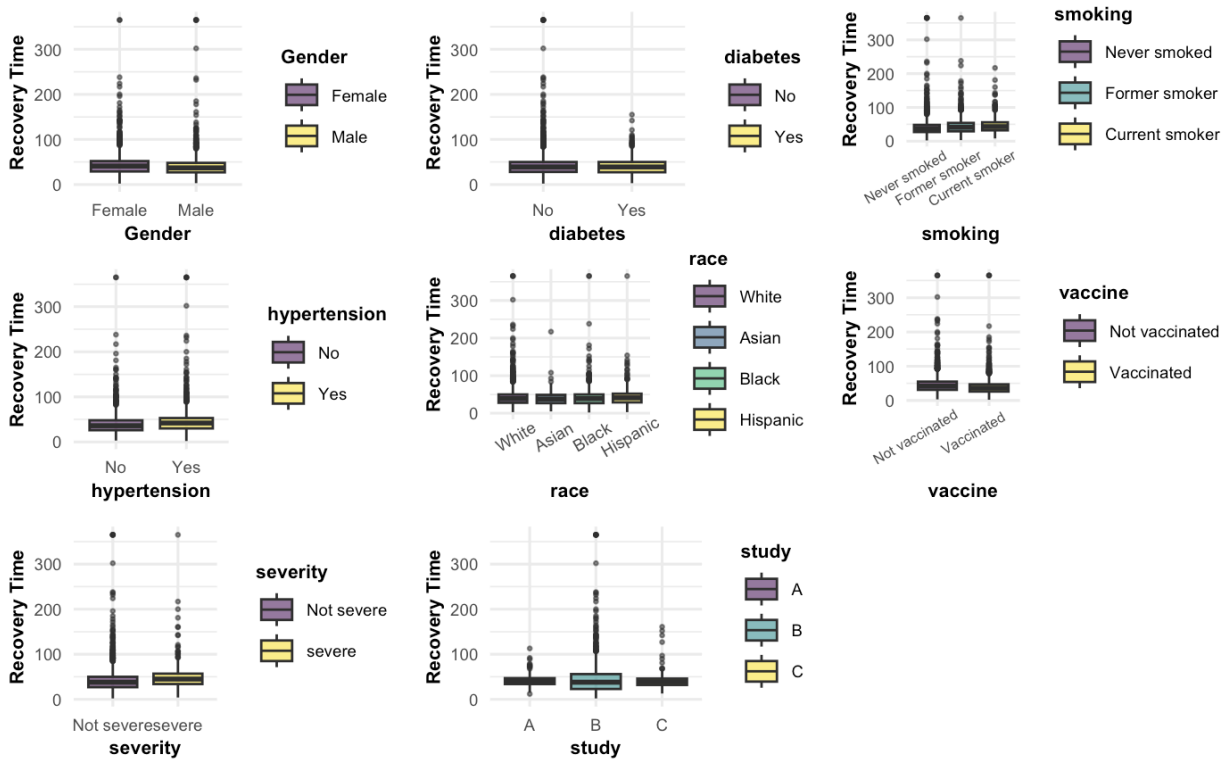

Figure 2: Distribution of Recovery Time by Categorical Predictors

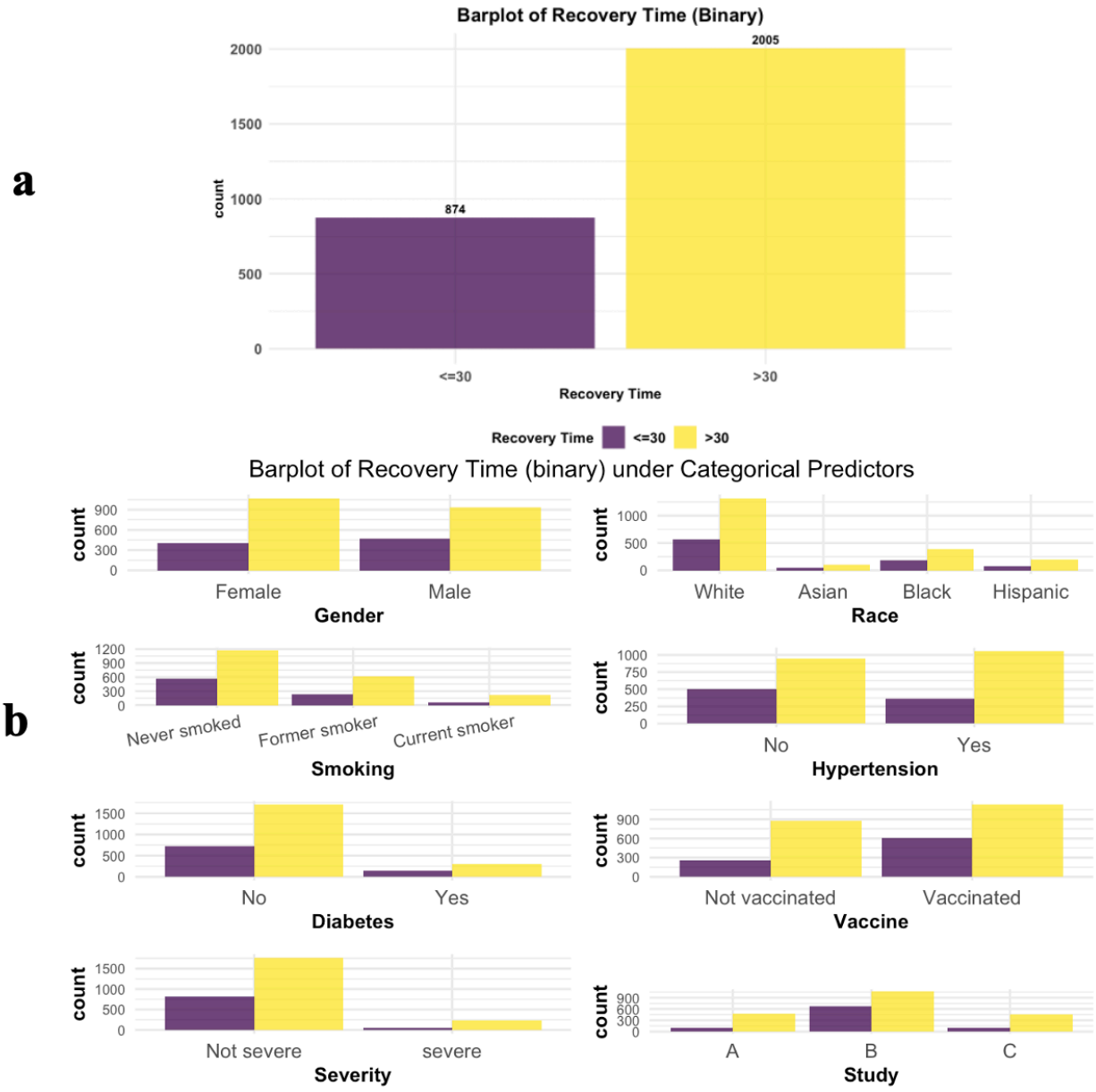

Figure 3: Distribution of Binary Response Variable Recovery Time

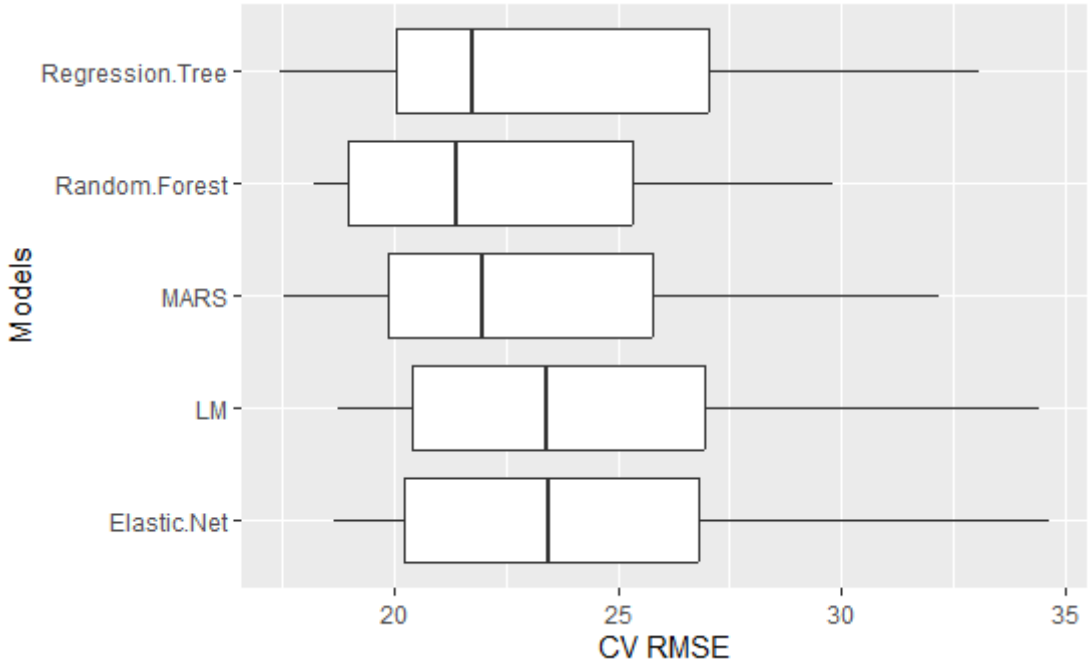

Figure 4: Cross Validation Error for Training Dataset under Five Models

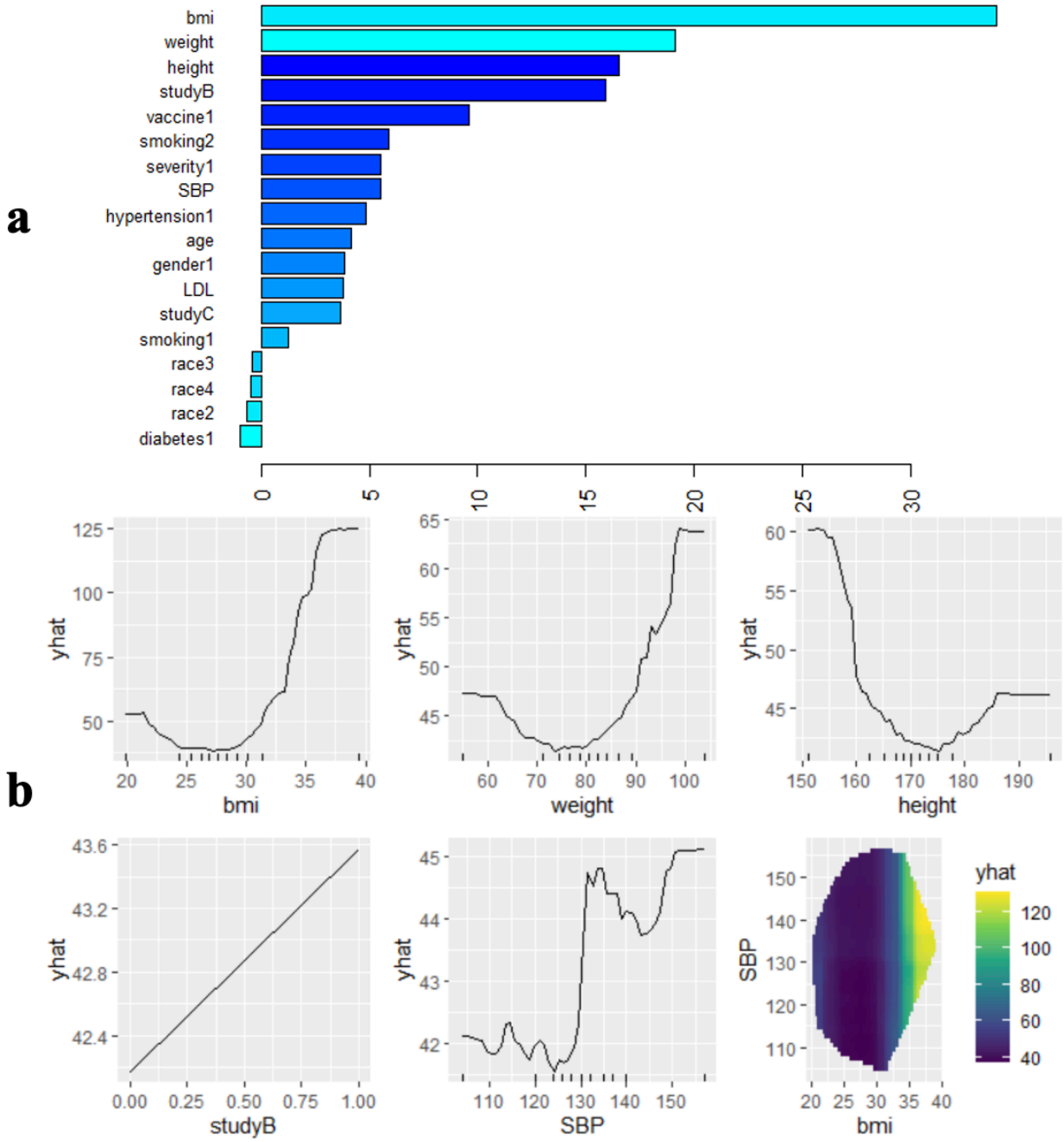

Figure 5: Variable Importance Plot and Partial Dependence Plot for Random Forest Model

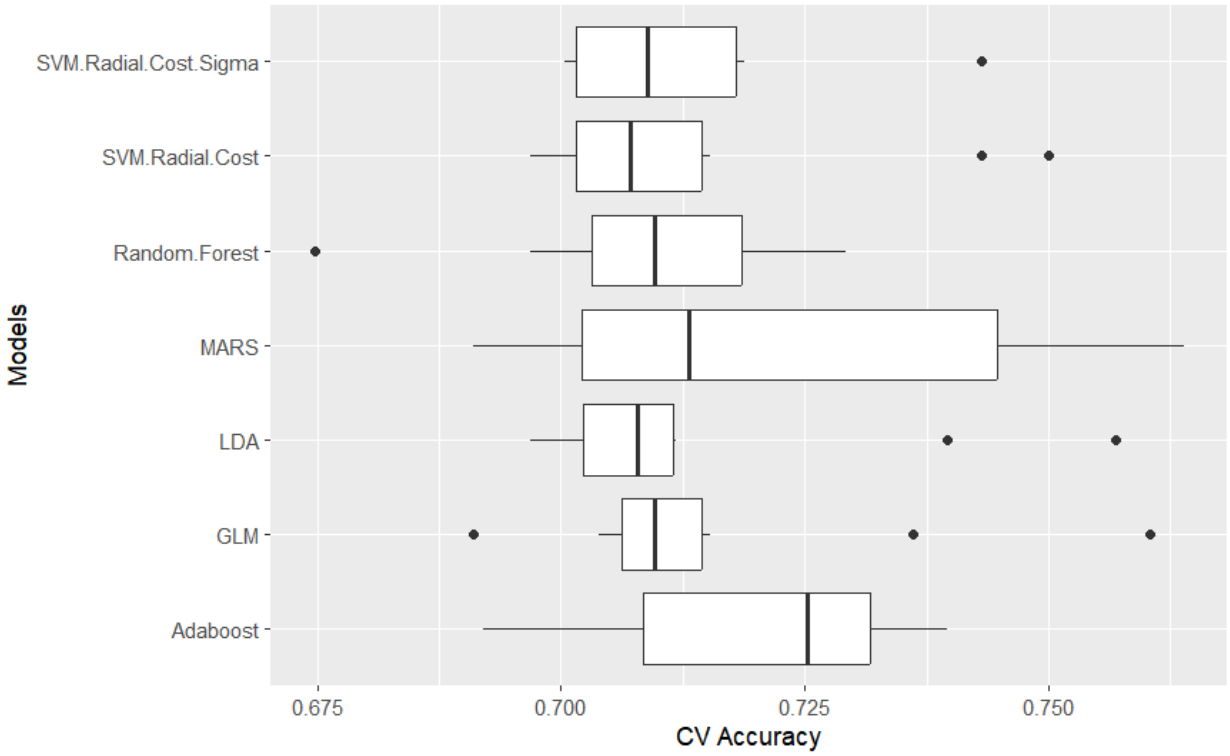

Figure 6: Cross Validation Accuracy for Training Dataset under Seven Models

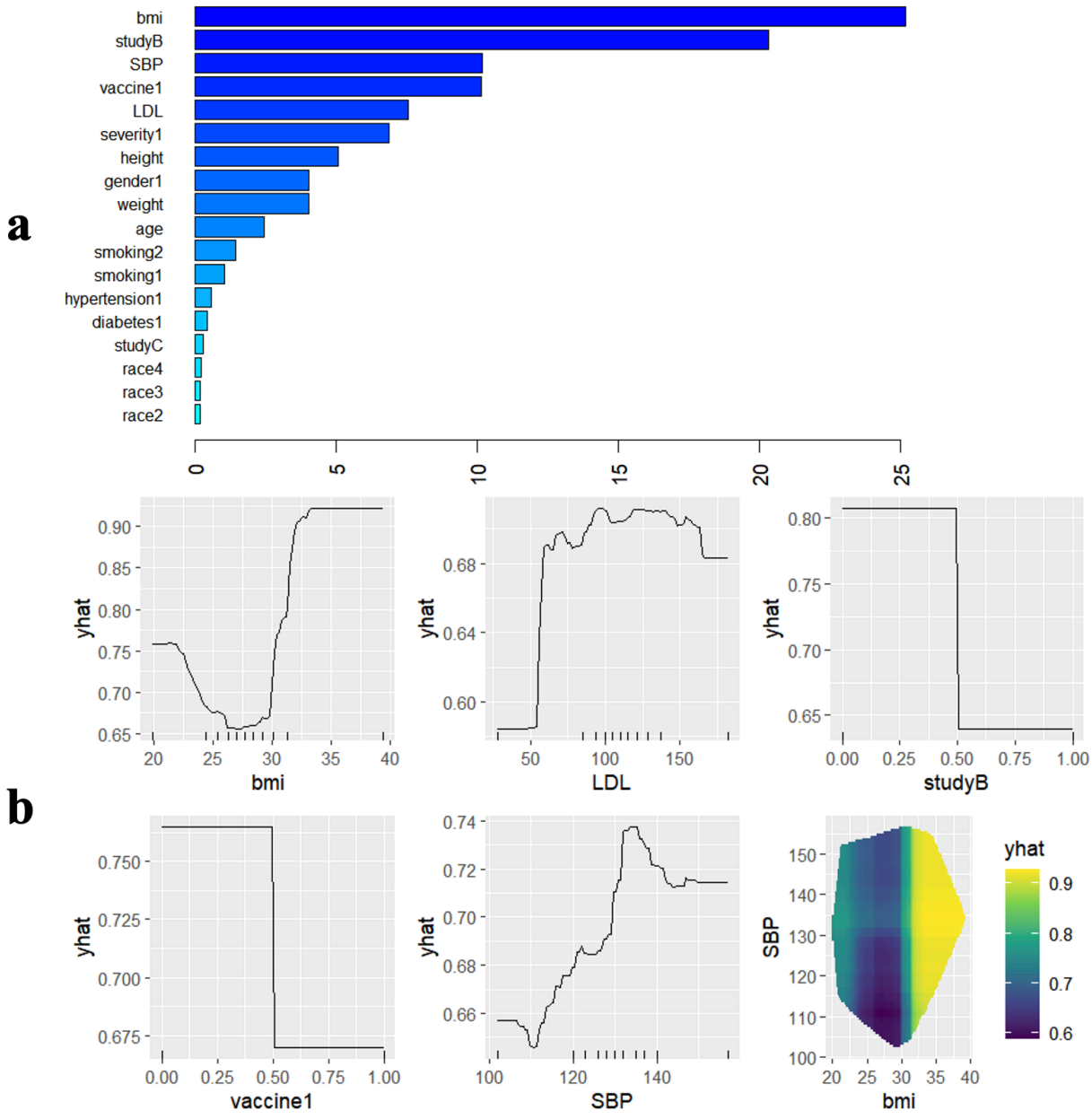

Figure 7: Variable Importance Plot and Partial Dependence Plot of AdaBoost Model