Predictive Analysis of Recovery Time from Covid-19

Chenyao Ni (cn2604) Wenjing Yang(wy2369) Yeque Sun(ys3508)

Introduction

The Covid-19 pandemic has affected millions of people worldwide, with many experiencing prolonged recovery times after contracting the virus. This project aims to develop a prediction model for Covid-19 recovery time and identify associated risk factors using data from three cohort studies. Additionally, a secondary analysis will present a binary outcome model for recovery time (>30 days vs. <= 30 days), providing a concise understanding of prolonged recovery and identifying at-risk individuals. The report will feature updated findings using merged data and more analysis techniques to enhance the accuracy and relevance of the results.

Variables Summary

- ID: Participant ID
- Gender: 1 = Male, 0 = Female
- Race/ethnicity: 1 = White, 2 = Asian, 3 = Black, 4 = Hispanic
- Smoking: Smoking status; 0 = Never smoked, 1 = Former smoker, 2 = Current smoker
- Height: Height (in centimeters)
- Weight: Weight (in kilograms)
- BMI: Body Mass Index; BMI = weight (in kilograms) / height (in meters) squared
- Hypertension: 0 = No, 1 = Yes
- Diabetes: 0 = No, 1 = Yes
- Systolic blood pressure (SBP): Systolic blood pressure (in mm/Hg)
- LDL cholesterol: LDL (low-density lipoprotein) cholesterol (in mg/dL)
- Vaccine: Vaccination status at the time of infection - 0 = Not vaccinated, 1 = Vaccinated
- Severity: Severity of COVID-19 infection-0 = Not severe, 1= Severe
- Study: The study (A/B/C) that the participant belongs to
- Time to recovery: Time from COVID-19 infection to recovery in days
- Length indicator: A binary variable indicating whether the participant's recovery time (in days) was longer than 30 days - 1 = Recovery time longer than 30 days, 0 = Recovery time no longer than 30 days

Exploratory analysis and data visualization

The scatterplots show that **SBP**, **LDL**, and **Age** are the continuous variables linearly correlated (small correlation coefficient though) to **Recovery time** while no apparent discrepancy is found when finding the association between **Recovery time** and the categorical variables of different levels. The correlation plot suggests that **SBP** is positively correlated to **Hypertension** while the correlation among **BMI**, **Weight**, and **Height** are foreseen. In short, there is no obvious collinearity among most of the covariates (none of them exceed the threshold of strong correlation 0.8).

The LDA partition plots show that the response variable is **Length_ind** and the predictor variables are **age**, **BMI**, **SBP**, and **LDL**, which are continuous variables. From the plots, we would like to visualize the classification performance of a linear discriminant analysis (LDA) model based on every combination of two variables. However, separation of classes is not effective in the plots.

Model Training (Primary analysis)

(Note: 5 times repeated 10 folds cross-validation is used across the training process)

- Ordinary Least Squares Regression (OLS)

While ordinary least squares regression assumes linear relationship, normality of residuals, homoscedasticity, and independence of residual error terms, it can be seen from the diagnostic plots that the assumptions of normally distributed residuals and homoscedasticity are violated. Therefore, regularization methods and nonlinear methods should be introduced.

- Elastic Net Regression & Partial Least Squares

Elastic net regression has no other assumption as it is a parameter regularization method to handle the multicollinearity of OLS. Through the process of cross-validation with alpha from 0 to 1 and a positive range for penalty coefficient, the resulting elastic net model gives an alpha at 1, which is equivalent to the Lasso model with a penalty coefficient of 0.135. As a dimension reduction method, partial least square regression assumes the independence between components and homogeneity and function nature of the system while making no assumption to the observed data. The partial least square model gives 14 components after cross-validation for best predictability within a tuning range from 1 to 18 since there are maximum 18 predictors in OLS.

For the non-parametric models below, there is no assumption of the underlying data.

- Generalized Additive Model (GAM)

For a generalized additive model, the tuning parameter is a decision for adding a penalty to each of the terms. The selected model after cross-validation has no penalty to each term, 6 out of 18 predictors get smooth terms while the rest of predictors are linearly related to the response variable.

- Multivariate Adaptive Regression Spline Model (MARS)

In the selected multivariate adaptive regression spline model after cross-validation, there are 3 out 18 predictors with 5 terms and a degree of 2 while the tuning parameters for terms range from 2 to 18 and degree from 1 to 3. Within the 3 predictors in the model, BMI is of most significance as it is included in 3 out of 5 terms.

- K-Nearest Neighbors Regression (KNN)

The selected KNN model has the tuning parameter k = 13 after cross-validation, with a tuning range of k from 5 to 15.

- Regression Tree (CART)

The selected model after cross-validation has a complexity parameter (cp) of 0.00778, which is chosen from a candidate grid of cp values ranging from exp(-6) to exp(-2) with 50 equally spaced values on a logarithmic scale.

- Random Forest

The mtry is the number of variables randomly sampled at each split when building a decision tree in the forest, and its optimal value depends on the size and complexity of the dataset, as well as the number of trees in the forest. A smaller mtry value can result in lower correlation between the trees in the forest in order to prevent overfitting. Here, the tuning range for mtry is set to 4 to 6, which is appropriate for a dataset with 18 predictors. The min.node.size parameter was also tuned between 3 and 8 to control the minimum size of terminal nodes in each tree, based on the number of observations in the training set. After cross-validation, the optimal number of randomly selected predictors is 6 and the minimal node size is 7.

- Generalized Boosted Regression (GBM)

The final generalized boosted regression model was created using 300 trees and 3 splits per tree, with a range of tree numbers from 100 to 1200 in increments of 100. The split depth of each tree was set to 1, 2, and 3. The learning rate was specified as 0.01 and 0.03, and each terminal node had at least 5 observations based on the relative size of the training set. (Based on Rule of Thumb: learning rate = max (0.01, 0.1*(min (1, nl/10000)), default terminal nodes as 10 observations for a training set size of 10000).

Model Training (Secondary analysis)

(Note: 10 folds cross-validation is used across the training process, ROC-AUC is selected as metric because of the imbalanced data)

- Penalized Logistic Regression

The best tune of the selected penalized logistic regression model within the tuning range from exp(-10) to exp(-2) is 0.000069. The lambda is close to zero, which indicates penalization term is not needed. From the results, the mixing proportion (alpha) is 1 so the model might have a lasso penalty term.

- Multivariate Adaptive Regression Spline Classifier (MARS)

The MARS model here has 9 terms within the range from 2 to 20 with 1 degree of freedom within the range from 1 to 3.

- Quadratic Discriminant Analysis (QDA)

Unlike Linear Discriminant Analysis (LDA), which assumes that the covariance matrix is the same for all classes, QDA allows for different covariance matrices for each class.

- Naive Bayes (NB)

The Native Bayes classifier can be used for nonparametric estimators. The Laplace Correction is implemented by the parameter fL, which is equal to 1. The bandwidth of the kernel density estimates is adjusted from 0.2 to 3. We also assume features in each class are independent in the NB model.

- Classification Tree

The reported model with highest cross-validation AUC gives a complexity parameter at 0.007, within the tuning range from exp(-6) to exp(-3).

- Adaptive Boosting

The resulting Adaptive Boosting model has 400 trees with 1 split performed on a tree with a tuning range of tree number from 100 to 1200 and split depth from 1 to 3, as the learning rate is specified as 0.01 and 0.03 and each terminal node has at least 5 observations based on the relative size of training set (Rule of Thumb used: learning rate = max(0.01, 0.1*(min(1, nl/10000)), default terminal nodes = 10 observations for size 10000).

- Support Vector Classifier (Linear kernel)

The support vector classifier with linear boundary has best cross-validation performance with a cost parameter at 0.49 within the tuning range from exp(-5) to exp(2).

- Support Vector Classifier (Radial kernel)

The support vector classifier with radial boundary has best cross-validation performance with a cost parameter at 0.19 within the tuning range from exp(-5) to exp(2), and gamma parameter (called sigma in kernlab package) at 0.0011 within the tuning range from exp(-8) to exp(-3), which controls the distance of the influence by a single training point.

Results

- Primary Analysis

The generalized boosted regression model is selected as the final model for its best training performance with the lowest cross-validation root mean square error (22.50) among all models. This boosted regression model has a test mean squared error of 492.57. The final model indicates that **BMI** is the most significant predictor, explaining over 61% of the reduction in the loss function with this set of predictors. **Study group**, **LDL**, and **weight** also play an essential role as predictors due to their mathematical relationship to **BMI**. A partial dependence plot shows that the recovery time is relatively short for individuals with a BMI between 23 and 30, whereas the obese patients with **BMI** greater than 30 tend to experience a remarkably long recovery period from Covid-19. **Vaccine**, another factor that draws the public's attention, does help to slightly shorten the recovery time as shown by the partial dependence plot while the recovery time increases mildly as **Age** increases for the elderly over 50. On the other hand, the variables such as systolic blood pressure (**SBP**), **Race**, and **Diabetes** have little influence on the recovery time from Covid-19.

- Secondary Analysis

The adaptive boosting model is selected as the final model for its best training performance with the highest mean cross-validation area under the ROC curve (0.71) among all models. This Adaboost model has a test accuracy of 0.72 and a kappa of 0.18. **Study group**, **BMI**, and **Vaccine** are believed to be the most important predictors, which account for more than 65% of the reduction to the loss function given this set of predictors. Partial dependence plot and

Individual Conditional Expectation plot show that the individuals with **BMI** between 23-30 have lower probability of getting long recovery time. Lower **SBP** level is associated with lower long recovery probability while the **LDL** level does not influence the long recovery probability significantly. **Vaccine** is also related to lower probability of having a long recovery time.

Conclusion

- Primary Analysis

Based on the generalized boosted regression model with the best performance in predictions, **BMI** is the single most important factor that influences long recovery time (as well as study group, LDL, and weight due to their mathematical relationship to BMI). Obese patients with a **BMI** over 32 will experience a remarkably longer recovery period compared to patients with a lower BMI. Other common risk factors, such as symptom severity and age, do have a positive correlation with the length of recovery time, while vaccination does shorten the recovery period. However, there is no evidence to suggest that these factors have an effect on the scenario of long recovery time as they only associate with a mild difference in recovery time.

- Secondary Analysis

Although the adaptive boosting model lists **Study group B** as the most important predictor, it can be largely attributed to the imbalanced distribution of data from the three cohort studies (See Appendix). For a new observation, this predictor would be no longer valid. **BMI** is still the most significant predictive variable as the figures show that the individuals with relatively normal **BMI** (23-30) have remarkably lower risk of experiencing long recovery time from covid, compared to either underweight or obese individuals (reduction of 17% and 27% respectively). In contrast to the results from primary analysis, **SBP** becomes a more influential variable in predicting whether an individual would experience long recovery time from covid though it might be given little weight in predicting the exact recovery time.