Midterm Report

Predicting Recovery Time of Patients with Covid-19

Yi Huang (yh3554)

2023-04-05

Contents

Abstract

This project applies varies regression methods to predict the recovery time of participants with Covid-19 to estimate the recovery time of these participants. The ultimate goal is to develop a prediction model for recovery time and identify important risk factors for long recovery time. The dataset is from a study that combines three existing cohort studies that have been tracking participants for several years.

1 Introduction

1.1 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.

1.2 Data Description

The dataset is a random sample of 2000 participants draw from "`recovery.RData` includes 16 variables and 2000 observations.
Description of each variable:
(1)Gender (gender): 1 = Male, 0 = Female; (2)Race/ethnicity (race): 1 = White, 2 = Asian, 3 = Black, 4 = Hispanic; (3)Smoking (smoking): Smoking status; 0 = Never smoked, 1 = Former smoker, 2 = Current smoker; (4)Height (height): Height (in centimeters); (5)Weight (weight): Weight (in kilograms); (6)BMI (bmi): Body Mass Index; BMI = weight (in kilograms) / height (in meters) squared; (7)Hypertension (hypertension): 0 = No, 1 = Yes; (8)Diabetes (diabetes): 0 = No, 1 = Yes; (9)Systolic blood pressure (SBP): Systolic blood pressure (in mm/Hg); (10)LDL cholesterol (LDL): LDL (low-density lipoprotein) cholesterol (in mg/dL); (11)Vaccination status at the time of infection (vaccine): 0 = Not vaccinated, 1 = Vaccinated; (12)Severity of COVID-19 infection (severity): 0 = Not severe, 1= Severe; (13)Study (study): The study (A/B/C) that the participant belongs to; (14) Age: age of participants; (15)Time to recovery (tt_recovery_time): Time from COVID-19 infection to recovery in days; (16) ID: unique id of each participant.
where time to recovery is our response variable; height, weight, bmi, Systolic blood pressure, LDL cholesterol, age are continuous predictors; the remaining are categorical predictors.

1.3 Data Cleaning

Package `tidyverse` and `dplyr` are used for data cleaning and wrangling. Variable `ID` is omitted before data wrangling. Table 1 includes the summary of all variables including 6 continuous and 9 categorical predictors and 1 continuous response. Based on the summary, there is no missing data. Then, factor all the categorical data before data partition. Apply `caret` package to split the data into 70% train data and 30% test data.

Table 1: **Summary of Dataset by Study Group**

Characteristic	**A**, N = 402	**B**, N = 1,209	**C**, N = 389
age	60.0 (57.0, 63.0)	60.0 (57.0, 63.0)	60.0 (57.0, 63.0)
height	170 (166, 173)	170 (166, 174)	170 (166, 174)
weight	79 (75, 84)	80 (76, 85)	80 (75, 85)
bmi	27.60 (25.60, 29.30)	27.70 (25.80, 29.50)	27.60 (25.60, 29.40)
SBP	131 (126, 136)	130 (125, 136)	129 (124, 135)
LDL	112 (98, 125)	110 (97, 123)	108 (96, 123)
recovery_time	40 (33, 47)	37 (23, 56)	39 (33, 45)
gender			

Characteristic	**A**, N = 402	**B**, N = 1,209	**C**, N = 389
0	211 (52%)	613 (51%)	191 (49%)
1	191 (48%)	596 (49%)	198 (51%)
race			
1	271 (67%)	779 (64%)	249 (64%)
2	21 (5.2%)	60 (5.0%)	15 (3.9%)
3	82 (20%)	257 (21%)	83 (21%)
4	28 (7.0%)	113 (9.3%)	42 (11%)
smoking			
0	242 (60%)	724 (60%)	249 (64%)
1	114 (28%)	357 (30%)	109 (28%)
2	46 (11%)	128 (11%)	31 (8.0%)
hypertension			
0	196 (49%)	629 (52%)	227 (58%)
1	206 (51%)	580 (48%)	162 (42%)
diabetes			
0	327 (81%)	1,017 (84%)	326 (84%)
1	75 (19%)	192 (16%)	63 (16%)
vaccine			
0	164 (41%)	471 (39%)	150 (39%)
1	238 (59%)	738 (61%)	239 (61%)
severity			
0	365 (91%)	1,091 (90%)	353 (91%)
1	37 (9.2%)	118 (9.8%)	36 (9.3%)

2 Exploratory analysis and data visualization:

In this section, I use several function to explore the data and identify the patterns by creating data visualizations. The function `ggpairs()` from `GGALLY` package allows to build a great scatterplot matrix, where the variable distribution is displayed on the diagonal, scatterplot of each pair of continuous variale on the left part of the figure, and pearson correlation is on the right figure. The function `ggplot()` from `ggplot2` generate density plot and boxlpots. In function `featureplot()` from `caret` can be used to plot the correlation of variables. In figure 1,

3 Model training

In this section, describe the models you used for predicting time to recovery from COVID-19. State the assumptions made by using the models. Provide a detailed description of the model training procedure and how you obtained the final model.

This project performs varies regression models on the dataset, including linear regression, K-Nearest Neighbors (KNN), ridge, lasso, lasso 1se, elastic net, partial least squares (pls), generalized additive model (gam), Multivariate Adaptive Regression (mars), random forest, and boosting using cross-validation results on train data. The model with the minimum mean of train RMSE from cross-validation results will be the best model. Finally, we estimate the test error for the best and second best models obtained from resampling cross-validation results to evaluate the test performance.

4 Results and discussion

In this section, report the final model that you built for predicting time to recovery from COVID-19. Interpret the results. Assess the model's training/test performance.

Training performance

Testing performance

5 Conclusions

In this section, summarize your findings from the model analysis and discuss the insights gained into predicting time to recovery from COVID-19.

References

CDC Covid Data Tracker. (2021), Centers for Disease Control and Prevention.

James, Gareth, e. a. (2021), An Introduction to Statistical Learning: With Applications in R., Springer.

Pradhan, A. and Olsson., P.-E. (2021), 'Sex differences in severity and mortality from covid- 19: are males more vulnerable?', Biology of sex differences 11(1), 53.

A First section of Appendix

B Second section of Appendix

The code used in this project are in the P8106_Code.pdf and P8106_Code.rmd files. Including data processing methods, modeling methods, tuning parameters selection, model comparison, etc. See Appendix