

Data science final guideline

Yi Huang, Yuchen Zhang, Shun Xie

2023-04-28

Contents

Final Project	1
Deliverables and deadlines	1
Dataset	1
Background	2
Data	2
Here is a description of each variable:	2
Report	3
Note	3
Introduction	3
Exploratory Data Analysis	4
Model training	4
Results	5
Conclusion	5
Appendix	5

Final Project

Deliverables and deadlines

Due: May 10 by 11:59pm
Each team is required to submit one report.

Dataset

Please merge the midterm project datasets from two team members, and keep only the unique observations.
If there are three team members, you can decide which two datasets to merge.

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.

Data

recovery.RData

The dataset in “recovery.RData” consists of 10000 participants. In your analysis, please draw a random sample of 2000 participants using the following R code:

```
set.seed([last four digits of your UNI])  
dat <- dat[sample(1:10000, 2000),]
```

The resulting dat object will contain a random sample of 2000 participants that you can use for your analysis.

Here is a description of each variable:

Variable Description

- ID (id) :Participant ID
- Gender (gender): 1 = Male, 0 = Female
- Race/ethnicity (race): 1 = White, 2 = Asian, 3 = Black, 4 = Hispanic
- Smoking (smoking): Smoking status; 0 = Never smoked, 1 = Former smoker, 2 = Current smoker
- Height (height): Height (in centimeters)
- Weight (weight): Weight (in kilograms)
- BMI (bmi): Body Mass Index; $BMI = \text{weight (in kilograms)} / \text{height (in meters)}^2$
- Hypertension (hypertension): 0 = No, 1 = Yes
- Diabetes (diabetes): 0 = No, 1 = Yes
- Systolic blood pressure (SBP): Systolic blood pressure (in mm/Hg)
- LDL cholesterol (LDL): LDL (low-density lipoprotein) cholesterol (in mg/dL)
- Vaccination status at the time of infection (vaccine): 0 = Not vaccinated, 1 = Vaccinated
- Severity of COVID-19 infection (severity): 0 = Not severe, 1 = Severe
- Study (study): The study (A/B/C) that the participant belongs to
- Time to recovery (tt_recovery_time): Time from COVID-19 infection to recovery in days

Report

Your report should be no more than five pages, excluding figures and tables. The total number of figures and tables should not exceed 8. You can submit an Appendix if you would like to include more details for modeling tuning.

For the primary analysis on time to recovery, you may revise your midterm report and incorporate new findings from the methods introduced in the second half of the semester.

As a secondary analysis, please consider time to recovery as a binary outcome (>30 days vs. ≤ 30 days) and develop a prediction model for this binary outcome.

Note

Similar to the midterm project report, your report should include the following sections: exploratory analysis and data visualization, model training, results, and conclusions.

```
library(caret)
library(corrplot)
library(tidyverse)
library(ggplot2)
library(dplyr)
library(ISLR)
library(glmnet)
library(mgcv)
library(nlme)
library(earth)
library(vip)
```

Introduction

```
#load data
load("data/recovery.RData")

#get data from 2 team members uni
set.seed(3554)
dat1 <- dat[sample(1:10000, 2000),]
set.seed(4437)
dat2 <- dat[sample(1:10000, 2000),]

#combine the data and discard duplicates
dat_temp <- rbind(dat1, dat2)
dat <- dat_temp[!duplicated(dat_temp$id),]

#dat is now the final data and it has the length as following:
length(dat$id)
```

```
## [1] 3593
```

Exploratory Data Analysis

Model training

```
# data manipulation
data =
  dat %>%
  mutate(study = factor(dat$study),
         gender = factor(dat$gender),
         hypertension = factor(dat$hypertension),
         diabetes = factor(dat$diabetes),
         vaccine = factor(dat$vaccine),
         severity = factor(dat$severity)) %>%
  dplyr::select(-id)
head(data)
```

```
##      age gender race smoking height weight  bmi hypertension diabetes SBP LDL
## 2989  63      1    1      0  174.8   83.6 27.4          1          0  137 108
## 620   60      1    1      0  170.0   80.7 27.9          0          0  128  80
## 9226  64      0    1      0  167.5   89.1 31.7          1          0  137 126
## 4098  69      0    1      2  173.0   72.5 24.2          1          0  135 149
## 8349  62      0    1      0  172.6   82.4 27.7          1          0  138 136
## 184   60      1    1      0  166.5   82.9 29.9          0          0  128 118
##      vaccine severity study recovery_time
## 2989        1         0    B             14
## 620         1         0    A             36
## 9226        0         0    C             50
## 4098        0         0    B             65
## 8349        1         0    C             34
## 184         1         0    A             32
```

```
#Split data into 80-20, using the third member's uni
set.seed(2337)
indexTrain <- createDataPartition(y=data$recovery_time, p=0.8, list=FALSE)

train_data <- data[indexTrain,]
test_data <- data[-indexTrain,]
```

primary analysis

Modeling Strategy

- Penalized logistic regression
- Generalized additive model (GAM)
- Multivariate adaptive regression splines (MARS)
- Linear discriminant analysis (LDA)
- Random Forest
- Boosting

Optimal Tuning Parameters

Resampling Results and Model Selection

secondary analysis

Modeling Strategy

- Penalized logistic regression
- Generalized additive model (GAM)
- Multivariate adaptive regression splines (MARS)
- Linear discriminant analysis (LDA)
- Classification and Regression Tree (CART)
- Random Forest
- Boosting
- Support Vector Machine (SVM with Linear and Radial Kernels)

Optimal Tuning Parameters

Resampling Results and Model Selection

Results

primary analysis

Interpretion

Variable Importance

model's training/test performance

secondary analysis

Interpretion

Variable Importance

model's training/test performance

Conclusion

Appendix