P8106 Data Science II Midterm Project Report: Predicting COVID-19 Recovery Time and Identifying Important Risk Factors

Sarah Forrest - sef2183

4/5/2023

Contents

Data

The dataset used in this study contains a variable for the time from COVID-19 infection to recovery (in days), which is the outcome of interest. It also contains 14 predictor variables, including demographic characteristics, personal characteristics, vital measurements, and disease status. The predictors are a mix of continuous and categorical variables. While the dataset consists of 10000 participants, a random sample of 2000 was used for analysis. Additionally, the sample was further split into training (70%) and test (30%) datasets using the `createDataPartition()` function. All data partitions were conducted using a seed set to my UNI number (2183) for reproducibility.

Exploratory Analysis and Data Visualization

Lattice plots were created using the `featurePlot()` function to visualize each potential predictor's association with the outcome, COVID-19 recovery time (see Figure 1 in the appendix). the following patterns were observed:

- Males appear to have a longer recovery time than females.
- White patients appear to have a shorter recovery time compared to all other races.
- Never smokers appear to have a shorter recovery time compared to former and current smokers.
- Patients with lower height values appear to have a slightly longer recovery time than patients with higher height values.
- Patients with lower weight values appear to have a slightly shorter recovery time than patients with higher weight values.
- A slight U-shaped association can be observed between patient bmi and recovery time. Patients with bmi values near the min and max in the dataset have longer recovery times than patients with bmi values in the middle.
- Patients with hypertension appear to have a longer recovery time compared to patients without hypertension.
- Patients with diabetes appear to have a longer recovery time compared to patients without diabetes.
- Vaccinated patients appear to have a shorter recovery time compared to unvaccinated patients.
- Patients with severe infections appear to have a longer recovery time compared to patients without severe infections.
- Patients in study B had a shorter recovery time compared to patients in studies A and C.
- The following variables do not appear to have a clear association with COVID-19 recovery time: patient age, SBP, and LDL cholesterol.

Model Training

Several different models were fit using all predictors to predict time to recovery from COVID-19. The `train()` function from the caret package was used to fit each model using either the training dataset or a matrix of all the predictors and a vector of the response variable, `recovery_time`, from the training dataset as inputs. Each model was also fit using cross-validation with 10 folds repeated 5-times. The `trainControl()` function was used to specify this cross-validation method, which was called on within the `train()` function using the trControl argument. Additionally, the method argument was used within the `train()` function to specify the type of model to fit. The resulting model object for each model contains the final model (finalModel) and information about the cross-validation performance. The `predict()` function from the caret package was also used to generate predictions for the test dataset using each final model that was trained using the training dataset. The root mean squared error (RMSE) between the predicted and actual recovery times on the test dataset was calculated in order to evaluate each model's performance and support model comparison.

1. Linear model A linear model that assumes a linear relationship between the predictor and response variables (linearity), and assumes that the errors are normally distributed (normality) and have constant variance (homoscedasticity),and that the observations are independent of each other (independence). The linear model is the most basic and assumes a linear relationship between the variables, while the other models allow for more flexible relationships. A method = "lm" argument was used within the `train()` function to specify a linear model fit.

2. Lasso model A lasso model is a linear regression model that adds a penalty term to the sum of absolute values of the coefficients to prevent overfitting, which can lead to sparse models by shrinking some coefficients to zero. It has the same assumptions as the linear model. A method = "glmnet" argument was used within the `train()` function to specify a lasso model fit. Additionally, a grid of tuning parameters for the model were set using the tuneGrid argument within the `train()` function. The grid contains 100 values of the lambda parameter, ranging from exp(-1) to exp(5) with equal intervals on the log scale. The alpha parameter is set to 1, indicating that we are only considering Lasso regularization. The `expand.grid()` function is used to generate all possible combinations of alpha and lambda in the grid for tuning.

3. Elastic net model An elastic net model is linear regression model that combines the penalties of the lasso and ridge regression methods to prevent overfitting, which can result in better prediction accuracy than either method alone. It has the same assumptions as the linear model. A method = "glmnet" argument was used within the `train()` function to specify a elastic net model fit. Additionally, a grid of tuning parameters for the model were set using the tuneGrid argument within the `train()` function. The grid includes 21 values of the alpha parameter, ranging from 0 to 1 with equal intervals. The lambda parameter is set to a sequence of 50 values, ranging from exp(2) to exp(-2) with equal intervals on the log scale. The expand.grid() function generates all possible combinations of alpha and lambda in the grid for tuning.

4. Partial least squares (PLS) model A PLS model is a model that seeks to find a low-dimensional representation of the predictor variables that explains the maximum variance in the response variable. It has the same assumptions as the linear model as well as a latent variable assumption, the assumption that the predictor variables are linearly related to the response variable via a set of underlying latent variables. A method = "pls" argument was used within the `train()` function to specify a PLS model fit. Additionally, a tuning parameter for the model were set using the tuneGrid argument within the `train()` function. The tuneGrid object includes a data frame with a single column ncomp that ranges from 1 to 17, representing the number of components used in the model. The preProcess argument is set to "center" and "scale", which means that the training data will be centered and scaled prior to model fitting.

5. Generalized additive model (GAM) model A GAM model is a model that extends the linear model by allowing for nonlinear relationships between the predictor and response variables, using flexible functions called splines. It has the same assumptions as the linear model. A method = "gam" argument was used within the `train()` function to specify a GAM fit. Additionally, two separate GAM models were fit: one GAM model using all predictors, and one GAM model with variable selection performed during model training. For the variable selection GAM model, a tuneGrid object was specified within the `train()` function. The tuneGrid object includes a data frame with two arguments: method is set to "GCV.Cp", which represents the generalized cross-validation with the Cp criterion for smoothing parameter selection, and select is set to TRUE, indicating that variable selection will be performed during model training.

6. Multivariate adaptive regression spline (MARS) model A MARS model is a model that uses piecewise linear or nonlinear functions to model the relationship between the predictor and response variables, which can capture complex nonlinear relationships. It has the same assumptions as the linear model. A method = "earth" argument was used within the `train()` function to specify a MARS model fit. Additionally, the `expand.grid()` function is used to generate a grid of tuning parameters. The mars_grid object includes two arguments: degree is set to 1, 2, and 3, representing the number of possible product hinge functions in a single term, and nprune is set to integers between 2 and 17, representing the upper bound on the number of terms in the model. The tuneGrid argument in the `train()` function uses the mars_grid object to specify the parameters for model tuning.

The final model was selected by comparing the median training RMSE for all models (see Table 1 and Figure 2 in the appendix). The lasso model had the highest mean and median training RMSE (i.e., worst

performance), and the MARS model had the lowest mean and median training RMSE (i.e., best performance). However, the MARS model only contained 4 terms with 2 of the predictors, so it was not an ideal model for comprehensively identifying important risk factors for recovery time, as so few risk factors were included. Therefore, the GAM model will all predictors–the second best model in terms of mean and median RMSE–was selected as the final model in this study.

Results

The final GAM model formula uses the `recovery_time` variable as the outcome, and all predictors in the dataset (see Table 2 in the appendix for the model results summary). Dummy variables were created for the patient race variable with "White" set as the reference category, the smoking variable with "Never Smoker" set as the reference category, and the study variable with "Study A" set as the reference category. Additionally, a smoothing function was applied to the numeric predictors (patient age, SBP, LDL cholesterol, BMI, height, and weight)

Model interpretation: Significant predictors of COVID-19 recovery time at the 5% level of significance include gender, smoking history, hypertension, vaccination status, infection severity, and study group. The GAM model also identified significant nonlinear relationships between BMI, weight, and height and predicted recovery time. On average:

- Being male is associated with a 5.3039-day shorter predicted recovery time than being female.
- Being a former smoker is associated with a 4.6506-day longer predicted recovery time than never smoking.
- Being a current smoker is associated with a 8.0189-day longer predicted recovery time than never smoking.
- having hypertension is associated with a 5.2189-day longer predicted recovery time than no hypertension diagnosis.
- Being vaccinated is associated with a 8.0292-day shorter predicted recovery time than no vaccination.
- severe COVID-19 infection is associated with a 9.5517-day longer predicted recovery time than non severe.
- Being in study B is associated with 4.2897-day longer predicted recovery time than being in study A.

Model performance: All of these factors together explain 43.8% of the deviance in COVID-19 recovery time. Additionally, the model's training error of 24.4 (see appendix) indicates that, on average, the model's predictions on the training data deviate from the actual values by 24.4 units. Meanwhile, the test error of 23.8 (see appendix) suggests that the model's performance on unseen data is slightly better than its performance on the training data. This indicates that the model generalizes well to new data.

Conclusion

The final GAM model using all predictor variables found that several factors were statistically significant in predicting time to recovery from COVID-19. On average, having a history of former or current smoking, having hypertension, and experiencing severe COVID-19 infection were all significantly associated with a longer predicted recovery time. Additionally, being in study B was associated with a longer recovery time compared to being in study A. On the other hand, being male and being vaccinated was associated with a shorter recovery time. Finally, BMI, height, and weight are also significantly associated with predicted COVID-19 recovery time. The model did not find significant associations with race or diabetes. These insights can be useful in predicting recovery time and informing clinical decisions in the management of COVID-19 patients.

Appendix

Figure 1. Lattice Plot

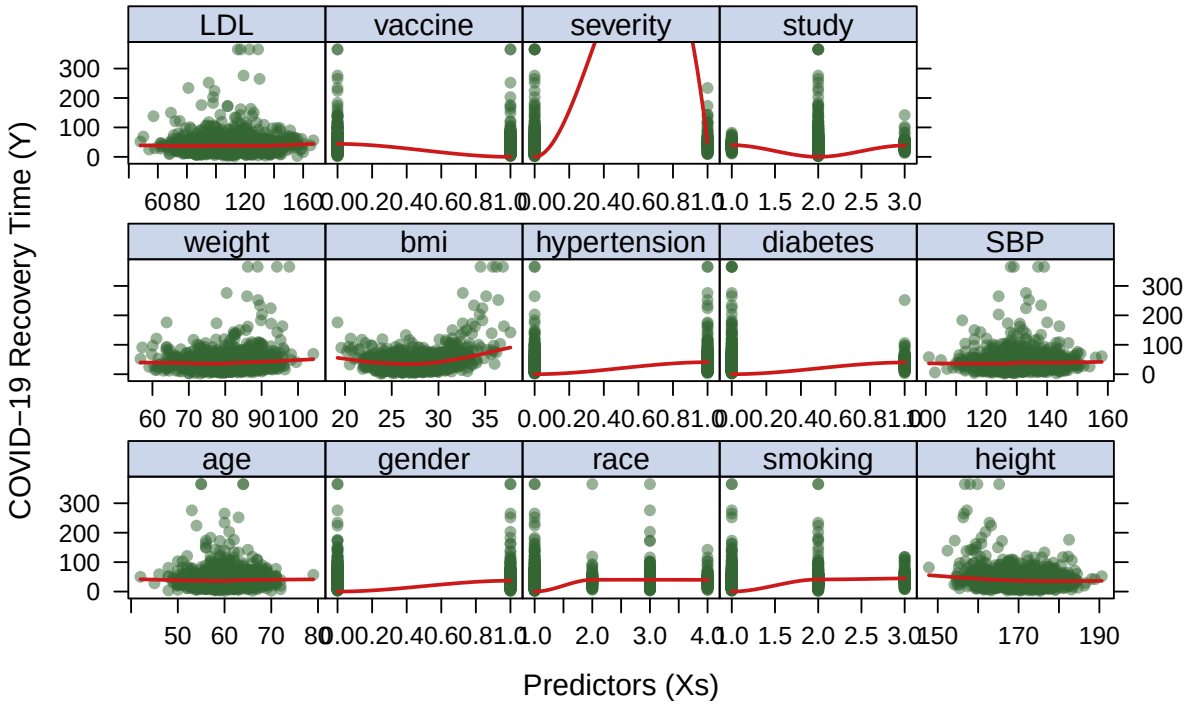

Table 1. Model Comparison

```
##
## Call:
## summary.resamples(object = resamp)
##
## Models: lm, lasso, enet, pls, gam_all, gam_select, mars
## Number of resamples: 50
##
## MAE
##                Min.  1st Qu.    Median      Mean  3rd Qu.      Max. NA's
## lm         12.73505 15.39825 16.12348 16.47142 17.75496 20.12710    0
## lasso      13.46116 15.56922 16.52594 16.76562 17.59255 21.58307    0
## enet       12.80380 14.94888 16.15967 16.06836 17.07309 20.37231    0
## pls        12.74936 15.38941 16.13187 16.47089 17.75006 20.13645    0
## gam_all    12.23868 14.72597 15.56555 15.59514 16.25176 18.68100    0
## gam_select 12.80144 14.36029 15.49947 15.61425 16.51267 19.37225    0
## mars       13.36260 14.03692 15.32847 15.39812 16.31531 19.08560    0
##
## RMSE
##                Min.  1st Qu.    Median      Mean  3rd Qu.      Max. NA's
## lm         16.82876 21.40034 23.77686 25.68307 29.70809 44.81248    0
## lasso      17.16745 22.15436 25.00758 27.85853 33.52794 52.01572    0
## enet       16.30816 20.86376 23.51353 26.01873 30.95194 48.57930    0
## pls        16.83601 21.40035 23.77552 25.68226 29.70413 44.82724    0
## gam_all    15.88988 21.15587 22.91025 24.37900 26.73050 39.29095    0
```

```
## gam_select 17.15290 20.86770 24.26783 24.58489 27.94064 37.21158    0
## mars       17.54908 20.46545 22.14328 23.64786 25.93674 37.85683    0
##
## Rsquared
##                  Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm         0.05285722 0.2198029 0.2767746 0.2885150 0.3387478 0.5731047    0
## lasso      0.01531994 0.1044821 0.1556263 0.1561066 0.1878872 0.3803743    0
## enet       0.04358373 0.2111552 0.2718273 0.2694211 0.3090849 0.5422976    0
## pls        0.05348801 0.2197665 0.2766038 0.2885046 0.3358747 0.5727813    0
## gam_all    0.10729883 0.2735852 0.3530194 0.3695334 0.4524725 0.6629995    0
## gam_select 0.09078299 0.2964500 0.3537030 0.3728204 0.4720036 0.6360189    0
## mars       0.04978937 0.2303193 0.3635106 0.3824594 0.5128183 0.7686433    0
```

Figure 2. Model Comparison Plot Using RMSE

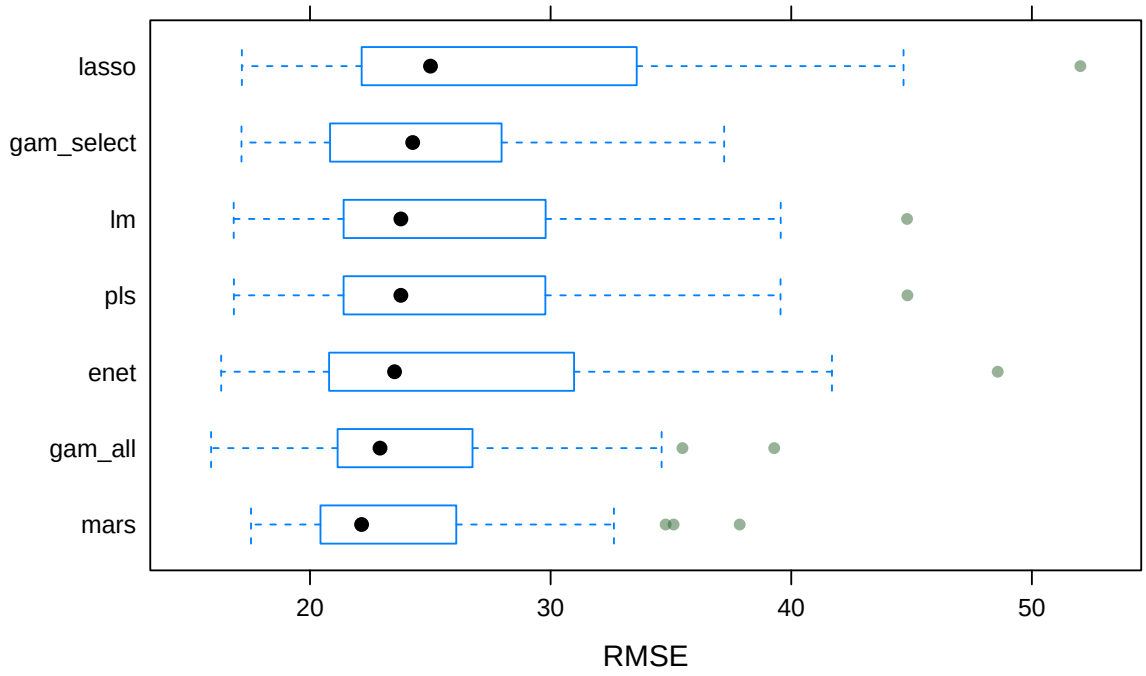

Table 2. Final GAM Model Summary

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ gender + race2 + race3 + race4 + smoking1 + smoking2 +
##     hypertension + diabetes + vaccine + severity + studyB + studyC +
##     s(age) + s(SBP) + s(LDL) + s(bmi) + s(height) + s(weight)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   42.6484     2.1195  20.122  < 2e-16 ***
## gender        -5.3039     1.2736  -4.165 3.31e-05 ***
```

```
## race2         -0.9167     2.6968  -0.340 0.733965
## race3          0.6748     1.6198   0.417 0.677042
## race4         -1.7393     2.2845  -0.761 0.446589
## smoking1       4.6506     1.4235   3.267 0.001114 **
## smoking2       8.0189     2.1284   3.768 0.000172 ***
## hypertension   5.2189     2.0964   2.490 0.012910 *
## diabetes      -2.1693     1.7875  -1.214 0.225124
## vaccine       -8.0292     1.2915  -6.217 6.72e-10 ***
## severity       9.5517     2.1691   4.404 1.15e-05 ***
## studyB         4.2897     1.6510   2.598 0.009470 **
## studyC        -1.5110     2.0152  -0.750 0.453522
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df      F  p-value
## s(age)    1.000  1.000  1.640  0.20058
## s(SBP)    1.000  1.000  0.034  0.85404
## s(LDL)    1.000  1.000  1.009  0.31524
## s(bmi)    8.795  8.985 66.151  < 2e-16 ***
## s(height) 7.741  8.605  4.054 5.81e-05 ***
## s(weight) 1.000  1.000  7.639  0.00579 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.425   Deviance explained = 43.8%
## GCV = 567.18  Scale est. = 553.61    n = 1402
```

Notes:

- "White" (`race` = 0) was set as the reference category for the `race` variable.
- "Never Smoker" (`smoking` = 0) was set as the reference category for the `smoking` variable.
- "Study A" (`study` = A) was set as the reference category for the `study` variable.
- An s() around the variable name indicates that a smoothing function was applied to the variable.
- An asterisk (*) next to the term indicates that the term was statistically significant at the 5% level of significance.

Final GAM model training error calculation (RMSE using the training data):

```
## [1] 24.37900 24.41081
```

Final GAM model test error calculation (RMSE using the test data):

```
## [1] 23.78453
```