P8106 Midterm Project

Ruihan Zhang

Rz2614

4/5/2023

**Introduction:**

A study was created to combine three current cohort studies that have been following participants for several years in order to better understand the variables that indicate recovery time from COVID-19 illness. The objective was to create a model for recovery time prediction and to find significant risk factors for prolonged recovery. The dataset in "recovery.RData" included 10000 participants. A random sample of 2000 participants were drawn to be analyzed.

**图示, 工程绘图

描述已自动生成Exploratory analysis and data visualization:**

图形用户界面

描述已自动生成Lattice plot and boxplot were used to explore the dataset and identify any patterns or relationships in the data. There were 14 predictors associated with the outcome. The association between each predictor and the outcome was shown in the plot. Patterns below were observed.

1. There was no obvious association between LDL cholesterol and recovery time. The almost straight line showed that patients with high or low LDL cholesterol have nearly the same recovery time.
2. For the vaccine variable, vaccinated patients were coded as 1, and unvaccinated patients were coded as 0. The negative association showed that vaccinated patients have a shorter recovery time than unvaccinated patients.
3. For the severity variable, severe patients were coded as 1, and not severe patients were coded as 0. Severe patients have a longer recovery time than not severe patients.
4. Patients in study B have a shorter recovery time than patients in study A and C who have almost the same recovery time.
5. There was no obvious association between weight and recovery time. The almost straight line showed that patients with high or low weight have nearly the same recovery time.
6. There was a positive association between BMI and recovery time. As BMI increases, the recovery time first decreases and then increases.
7. For the hypertension variable, patients with hypertension were coded as 1, and patients without hypertension was coded as 0. Patients with hypertension have a longer recovery time than patients without hypertension.
8. For the diabetes variable, patients with diabetes were coded as 1, and patients without diabetes was coded as 0. Patients with diabetes have a longer recovery time than patients without diabetes.
9. There was no obvious association between SBP and recovery time. The almost straight line showed that patients with any systolic blood pressure level have nearly the same recovery time.
10. There was no obvious association between age and recovery time. The almost straight line showed that patients with any age level have nearly the same recovery time.
11. For the gender variable, male patients were coded as 1, and female patients was coded as 0. Female patients have a longer recovery time than male patients.
12. For the race variable, white patients were coded as 1, Asian patients were coded as 2, black patients were coded as 3, and Hispanic patients was coded as 4. White patients have a relatively short recovery time compared with other races. Asian patients, black patients and Hispanic patients have almost the same recovery time.
13. For the smoking variable, patients never smoking were coded as 0, patients smoking formerly were coded as 1, current smoking patients were coded as 2. Patients never smoking have a shorter recovery time than others. Patients smoking formerly have a slightly shorter recovery time than current smoking patients.
14. The association between height and recovery time was slightly negative, which meant that as height increases the recovery time decreases slightly.

**Model training:**

Six models were used in this project, which are the linear model, LASSO model, GAM model, MARS model, elastic net model, and PLS model.

The linear model is a simple approach for supervised learning. It assumes that the dependence of outcome on predictors is linear, and that the errors have the constant variance and normal distribution. This linear model was fitted by train function with 10 folds cross validation for 5 times, which was done by the trainControl function. The method was linear regression. The model was summarized to get statistical details. The RMSE was calculated by using the predicted and actual value of recovery time.

The Lasso model uses the penalty term which forces some of the coefficient estimates to be zero to prevent overfitting. It performs variable selection and yields sparse models. It assumes that the dependence of outcome on predictors is linear, and that the errors have the constant variance and normal distribution. This Lasso model was fitted by train function with 10 folds cross validation for 5 times. The method was glmnet within the train function. The argument tuneGrid was used to set tuning parameters with length of 100 and range from exp(-1) to exp(5). The alpha parameter one indicated this was a Lasso model. The expand.grid function was used to get all possible combinations of lambda and alpha. The model was summarized to get statistical details. The RMSE was calculated by using the predicted and actual value of recovery time. The model was summarized to get statistical details. The RMSE was calculated by using the predicted and actual value of recovery time.

图表

中度可信度描述已自动生成

图表

描述已自动生成The elastic net model uses the ridge-type penalty to get the effective regularization and the lasso penalty to get the feature selection. It’s a more effective way to deal with groups of highly correlated predictors than the Lasso model or the ridge model. It assumes that the dependence of outcome on predictors is linear, and that the errors have the constant variance and normal distribution. This model was fitted by train function with 10 folds cross validation for 5 times, which was done by the trainControl function. The method was glmnet within the train function. There were 21 values of alpha parameter from 0 to 1. There were 50 values of lambda parameter from exp(-2) to exp(2). The expand.grid function was used to get all possible combinations of lambda and alpha. The model was summarized to get statistical details. The RMSE was calculated by using the predicted and actual value of recovery time.

图表, 折线图

描述已自动生成The PLS model makes use of the response variable to identify new features that not only approximate the old features well, but also are related to the response variable. It assumes that the dependence of outcome on predictors is linear, and that the errors have the constant variance and normal distribution. It also assumes that the predictor has a linear relationship with the response by a set of underlying latent variables. This model was fitted by train function with 10 folds cross validation for 5 times, which was done by the trainControl function. The method was pls within the train function. The model was summarized to get statistical details. The RMSE was calculated by using the predicted and actual value of recovery time. The tuneGrid used a data frame with one column ncomp ranging from 1 to 15, specifying the number of components in this model. The preProcess was set to “center” and “scale” to center and scale the training data before the model fitting. The model was summarized to get statistical details. The RMSE was calculated by using the predicted and actual value of recovery time.

图表, 折线图

描述已自动生成The MARS model creates a piecewise linear model by using the algorithm to select cut points automatically. There are two tuning parameters, the degree of features and the number of terms. It can show complex nonlinear relationships. It assumes that the dependence of outcome on predictors is linear, and that the errors have the constant variance and normal distribution. This model was fitted by train function with 10 folds cross validation for 5 times, which was done by the trainControl function. The method was earth within the train function. The degree was from 1 to 3, and the nprune was from 2 to 15 within the expand.grid function to produce a grid of tuning parameters. The model was summarized to get statistical details. The RMSE was calculated by using the predicted and actual value of recovery time.

The GAM model allows for flexible nonlinearities in several variables but retains the additive structure of linear models. It uses splines which are flexible functions. It assumes that the dependence of outcome on predictors is linear, and that the errors have the constant variance and normal distribution. This model was fitted by train function with 10 folds cross validation for 5 times, which was done by the trainControl function. The method was gam within the train function. The model was summarized to get statistical details. The RMSE was calculated by using the predicted and actual value of recovery time.

**图表, 折线图

描述已自动生成Model Comparison:**

图表, 箱线图

描述已自动生成文本

描述已自动生成The RMSE of each model was calculated. The model with the smallest mean RMSE should be the model with the best performance.

From the boxplot and the table above, the GAM model had the lowest mean RMSE and median RMSE, so the final model to predict the recovery time from COVID-19 was the GAM model.

**Results:**

The final model for predicting time to recovery from COVID-19 was the GAM model. The response variable was recovery time. The variables were gender, race2, race3, race4, smoking1, smoking2, hypertension, diabetes, vaccine, severity, study B, and study C. The reference category for race variable was white. The reference category for smoking variable was never smoked. The reference category for study variable was study A. An asterisk (\*) meant that this predictor is statistically significant. The statistically significant predictors were gender, smoking1, smoking2, vaccine, severity, study B, and study C. The smoothing terms were age, SBP, LDL, bmi, height and weight, and bmi, height and weight are significant.

From the final model, male patients have 5.825 days of recovery time shorter than female patients, Asian patients have 4.7591 days of recovery time shorter than white patients, black patients have 2.3053 days of recovery time shorter than white patients, Hispanic patients have 2.2304 days of recovery time shorter than white patients, former smoking patients have 3.5753 days of recovery time longer than patients never smoking, current smoking patients have 7.5368 days of recovery time longer than patients never smoking, patients with hypertension have 1.7696 days of recovery time longer than patients without hypertension, patients with diabetes have 2.1977 days of recovery time longer than patients without diabetes, patients vaccinated have 8.1433 days of recovery time shorter than patients unvaccinated, severe patients have 8.9118 days of recovery time longer than not severe patients, patients in the study B have 4.3804 days of recovery time longer than patients in the study A, and patients in the study C have 0.1369 days of recovery time longer than patients in the study A.

For the model performance, the RMSE from the training data was about 24. It indicated that the training data had the deviation about 24 units from the actual value. The RMSE from the testing data is about 19. It indicated that the testing data had the deviation about 19 units from the actual value. It also indicated that the testing data predicted better than the training data.

**Conclusion:**

By comparing the linear model, the Lasso model, the elastic net model, the GAM model, the PLS model, and the MARS model, the GAM model with the lowest mean RMSE performed better than other models. The significant variables are gender, smoking, vaccine, severity, study, BMI, height and weight. Thus, patients should take vaccines, give up smoking and do physical exercises in order to be recovered from COVID-19 more quickly.

**图示, 工程绘图

描述已自动生成**图形用户界面

描述已自动生成**Appendix**:

图表, 折线图

描述已自动生成

图表

描述已自动生成图表

中度可信度描述已自动生成

图表, 折线图

描述已自动生成

文本

描述已自动生成**图表, 折线图

描述已自动生成**

图表, 箱线图

描述已自动生成

图片包含 图示

描述已自动生成