P8106 Final Project

Zekai Jin, Zizhao Lin, Jiawen Zhao
05/08/2023

Introduction

COVID-19, caused by SARS-CoV-2 infection, can lead to a diverse range of clinical manifestations. After infection, most people will have mild to moderate symptoms, and will recover in 2 weeks. However, some people still suffer from prolonged illness, Thus, studying its relationship with other variables will be helpful for improving the recovery rate.

Data Description and exploratory Analysis

The dataset consists of 3576 observations of 14 predictors, 5 of which are binary, 3 are categorical, and 6 are continuous. The outcome is recovery time (in days), which is treated as a continuous variable. We also set recovery time as a binary variable taking values of 'f' and 's' for classification. Here, 'f' represents recovering within with equal 30 days and 's' represents recovering longer than 30 days. A summary of the variables of the dataset can be found in Supplementary files (Sup.1).

For continuous predictors, a pairwise scatter plot is performed to study their marginal relationship with the outcome (Fig.1). According to the plot, all predictors excluding BMI seem to have no relationship to the outcome, yielding a straight line for all values. BMI relates to the outcome through a “U” like function such that observations with moderate BMI have less recovery time. A correlation plot is also performed (Sup.2) and no obvious co-linearity is detected.

For categorical and binary predictors, a density plot was performed (Fig.2). Similarly, no obvious difference is seen except for study B. The distribution of recovery time in study B seems to have lower mean and higher variance compared to A and C.

Model Training

Our objection consists of 2 parts: regression on recovery time and classification on whether the recovery time is long or short. To address different possible underlying structures of the data, a variety of models were trained. For regression, Linear Model (LM), Elastic Net (Enet), Generalized Linear Model (GLM), Partial Least Squares (PLS), Generalized Additive Model (GAM), Multiple Adaptive Regression Splines (MARS) and some regression-tree based models were used. For classification, Generalized Linear Model (GLM), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and classification tree with boosting were compared. Before training, the data was randomly split into 80% of training set (n=2845) and 20% of testing set (n=731). The training set was used for model training, parameter tuning and model comparison by 10-fold cross validation (CV). The test set was used for estimating the final prediction accuracy. The criteria for model comparison are Root of Mean Squared Error (RMSE) for regression and Accuracy for classification. The calculation was done using caret (6.0.93) in R (4.2.2).

Description of Regression Models

Linear Model LM suggests a linear Relationship between response and predictors:

$$\vec{y} = X\,\vec{\beta} + \vec{\epsilon}$$

Here, $\vec{y}$ is the outcome vector, X is the data matrix and $\vec{\beta}$ is the coefficient vector. $\vec{\epsilon}$ is the error term such that $\epsilon_i \sim iid\ N(0, \sigma^2)$. The coefficient is obtained by minimizing the loss:

$$\left|\left|\vec{y} - X\,\vec{\beta}\right|\right|_2$$

Diagnosis plot suggests that the residual is not normal, which violates the model assumption (Sup.3). Thus, LM is not a good fit of the data.

Elastic Net Enet introduces L1 and L2 penalty to LM, which restricts the number and size of the coefficients by minimizing the following loss function:

$$\left|\left|\vec{y} - X\,\vec{\beta}\right|\right|_2^2 + \lambda\left[(1-\alpha)\left|\left|\beta\right|\right|_2^2 + \alpha\left|\left|\beta\right|\right|_1\right]$$

The parameter λ controls the strength of penalty and α controls the balance between L1 and L2 penalty. Both parameters are tuned by 10-fold CV (Fig). The result suggests that no penalty ($\lambda = 0$) yields the best prediction error, which leads to a linear model (Sup.4).

Generalized Linear Model Because the response variable is time for recovery, one may argue that it may not follow normal distribution as LM suggests. Instead, it should follow gamma distribution since recovery time is always positive and right skewed. This yields the following GLM:

$$\log(\vec{u}) = X\,\vec{\beta}$$

Where y_i follows a Gamma distribution with expectation u_i. The coefficient is obtained by maximum likelihood calculated by software.

Partial Least Squares PLS aims to reduce the input dimension by finding the best linear combination of inputs that maximizes the following:

$$Var(X\,\vec{\phi})Corr^2(\vec{y}, X\,\vec{\phi})$$

$$w.r.t.\left|\left|\vec{\phi}\right|\right|_2 = 1, \vec{\phi}^T Corr(X)\vec{\phi} = 0$$

The number of components is selected by the same 10-fold CV as before. The result suggests that 9 components as appropriate (Sup.5).

Generalized Additive Model GAM smooths the relationship between predictors and outcome, while holding the additive structure:

$$g(\vec{u}) = \beta_0 + \sum_{i=1}^{p} f_i(\vec{x_i})$$

The corresponding smooth functions f_i and coefficients are automatically selected by algorithm. In GAM, feature selection improved the RMSE by 0.05 compared to the full model (Sup.6).

Multiple Adaptive Regression Splines MARS is another model for non-linearity by creating a piecewise linear model with interaction term included. The cut points and slopes are selected by software. Two parameters are tuned: the maximum degree of interaction term and the number of terms. The result suggests that a degree 3 model with 6 terms performed the best (Sup.7).

Regression Tree Based on one predictor at each split, the decision tree splits the sample into one of its two branches, until it reaches the leaf node. The predicted value of the outcome is the mean value of all samples within the same leaf. The best result we can get is the regression tree when the complex number is 0.003 (Sup.8).

Random Forest RF is an ensemble method that combines the result of different trees. This increases stability and potentially Accuracy. The performance of the model did not differ by minimum node size and reached its maximum when 10 predictors were considered (Sup.9).

Boosting This approach grows a sequence of trees that each one is an update to the previous one. The parameters are learning rate, number of trees and number of splits in each tree and they are tuned using CV (Sup.10).

Description of Classification Models

Generalized Linear Model Because the response variable is now binary, one of the most straightforward methods is the logistic regression with link function:

$$\text{logit}(\vec{u}) = X\,\vec{\beta}$$

Where y_i follows a Bernoulli distribution with expectation u_i. The coefficient is optimized by software.

Linear Discriminant Analysis LDA assumes that each of the two classes follows a multivariate gaussian distribution. The shape of the distribution is estimated using training samples. The predicted class of a new data point will be the class with highest posterior density:

$$argmax_k(x^T u_k - u_k^T u_k + \log(\pi_k)$$

Here, the prior for each class is set to equal.

Support Vector Machine SVM suggests the two classes to be separable by a hyperplane. By choosing different kernel functions, the separation can be performed both in the original space (linear kernel) or in the transformed space (radial and polynomial kernel). Each of the kernels requires different parameters (Sup.11).

Classification Tree and Boosting These methods are similar to those in regression. Compared to the tree-based method for regression, in classification, the leaf nodes can only take on of the two values: long or short (Sup.12). Boosting is also performed to increase stability and an approximate of 2000 boosting iterations with learning rate 0.005, max tree depth of 2 is preferred (Sup.13).

Results

In total, 9 models were compared for regression and 5 models for classification. Model comparison is done using the same 10-fold CV dataset. The result is shown in Fig.5. In both tasks, boosted tree models outperformed other models, yielding the smallest cross-validated RMSE or prediction error rate. These two models will be selected as our final model and will be sent for further analysis.

In order to estimate the performance of the final models, compare the predicted value and true value for the test set. The resulting RMSE for regression is 22.34, meaning that on average, the predicted recovery time will differ the true value by 22.3 days. The prediction Accuracy for classification is 72.64%, meaning that on average, the correct predictions will consist of 73% of all predictions.

The predictor importance on regression is shown in Fig.5 and Fig.6. For regression, the most important variable is BMI, followed by study. Other predictors seem to have much less importance. For classification, the most important variable becomes study, followed by BMI. However, the other variables are also estimated to be important, which differs from regression. According to EDA, there are outliers in the dataset who have extremely long recovery time. These outliers, that produce high loss in regression, could have altered the importance of the predictors.

Conclusion

Based on the results of our final model, our observation in EDA that recovery time is mostly correlated with BMI and study is supported. However, boosting on trees is mainly a black-box model. Thus, we cannot only infer from the model on how important the predictors are, but not how they contribute to the outcome, both independently and interactively.

From the model comparison plots, models considering non-linearity performed slightly better than linear models. This suggests a subtle non-linearity of the dataset, possibly the U-like curve of BMI.

In conclusion, our dataset of 3576 observations and 14 predictors suggests a non-linear relationship of the recovery time and other predictors, mainly BMI of the subject and the study they belong to. For both regression and classification, boosted tree models outperformed other models and the final test error is 22.34 (RMSE) for regression and 72.64% (Accuracy) for classification.

Fig.1 Scatter plot of response variable vs. continuous predictors

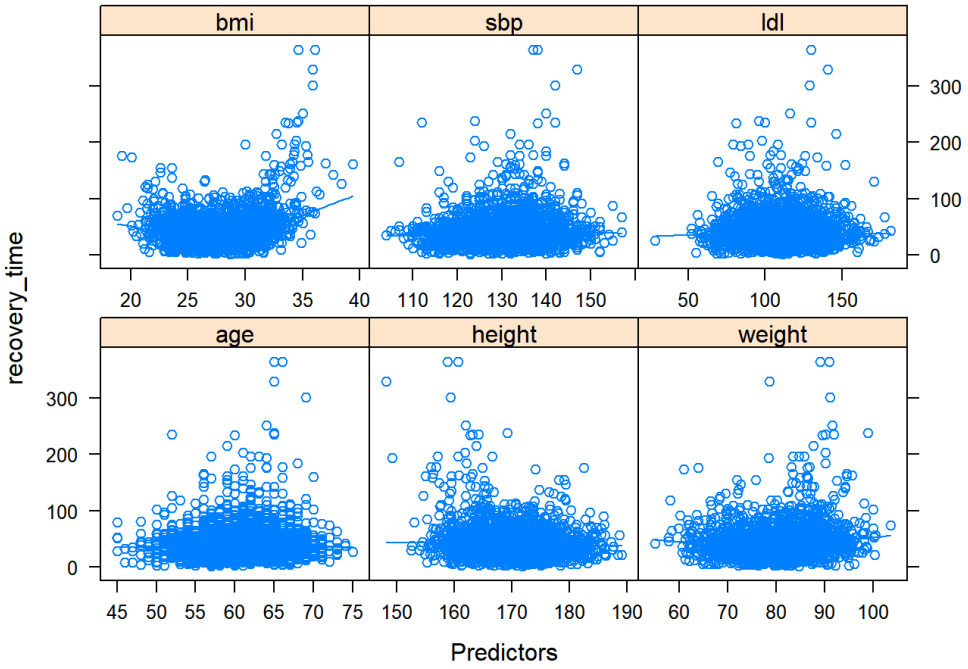

Fig.2 Density plot of response variable among different categorical variables

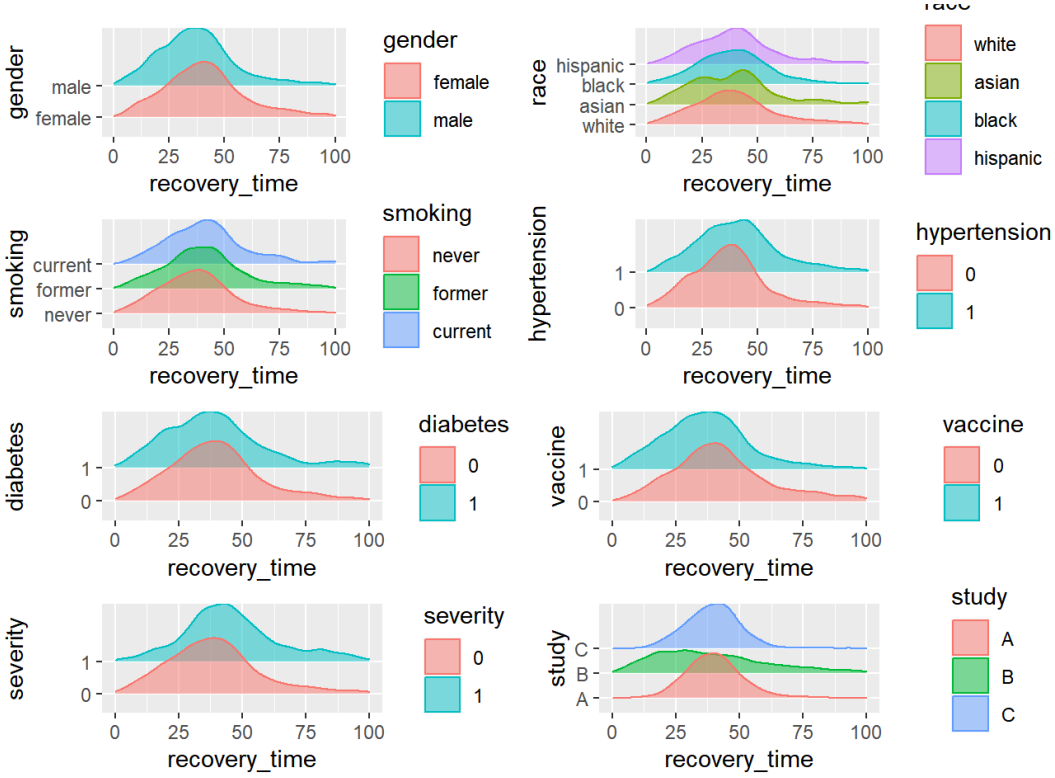

Fig.3 Comparison of different regression models

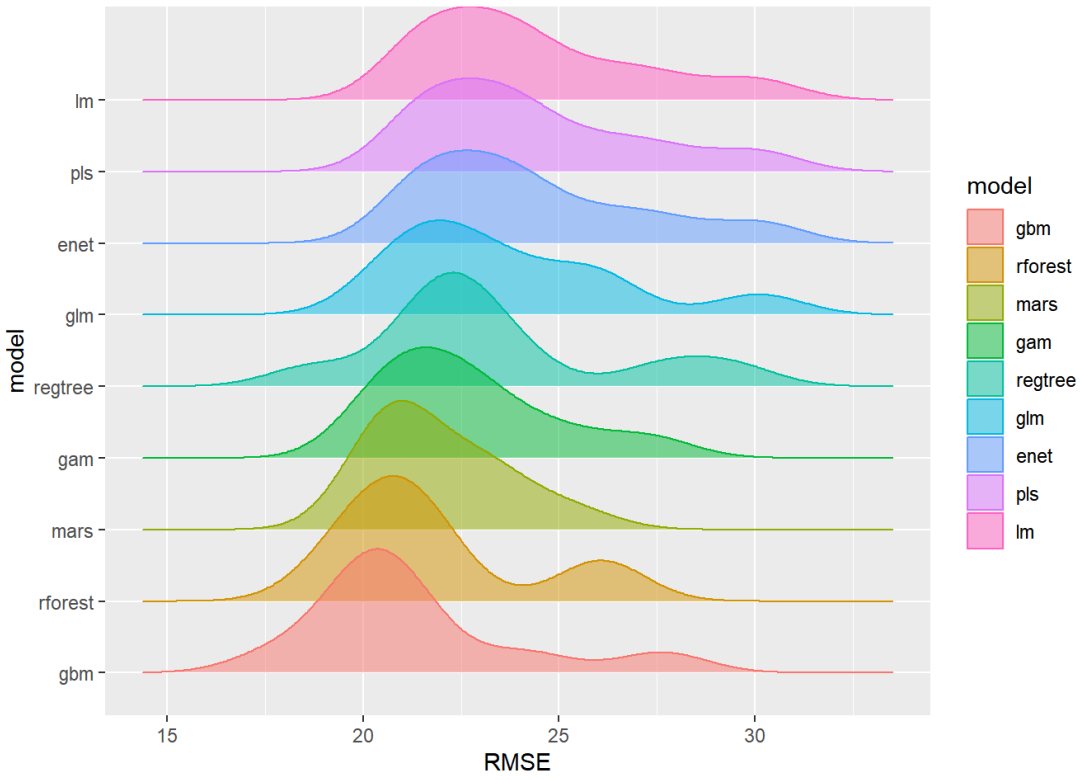

Fig.4 Comparison of Accuracy for different classification models

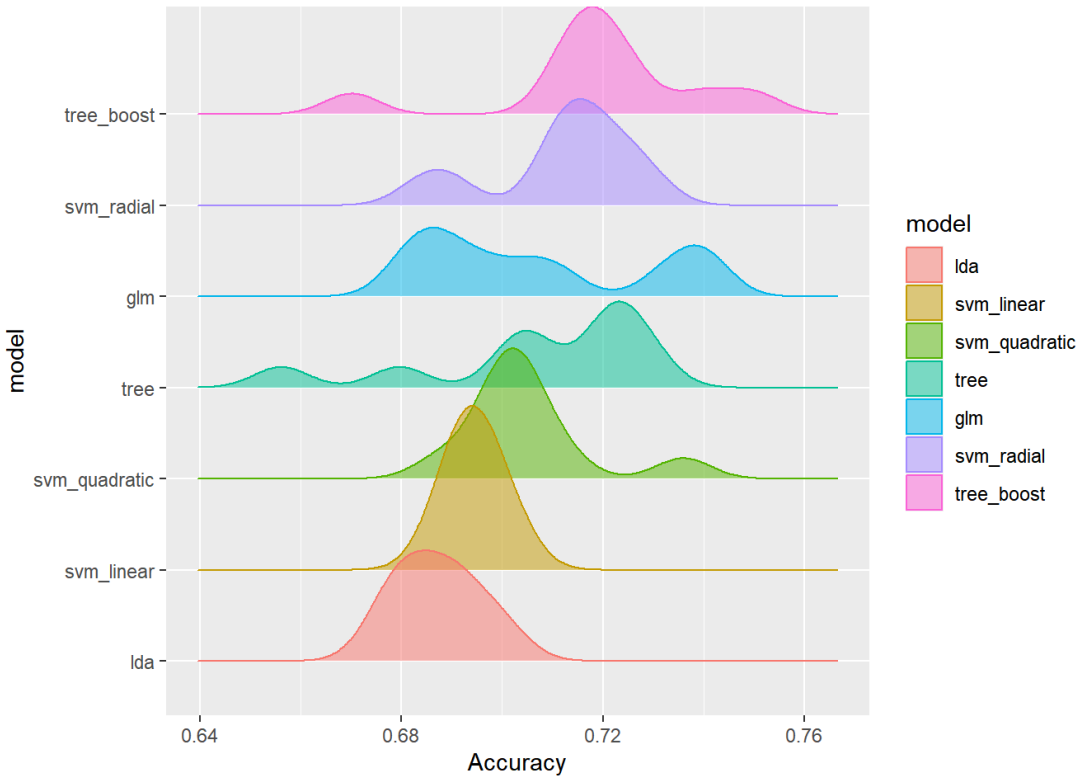

Fig.5 Variable importance plot of the Boosting regression tree

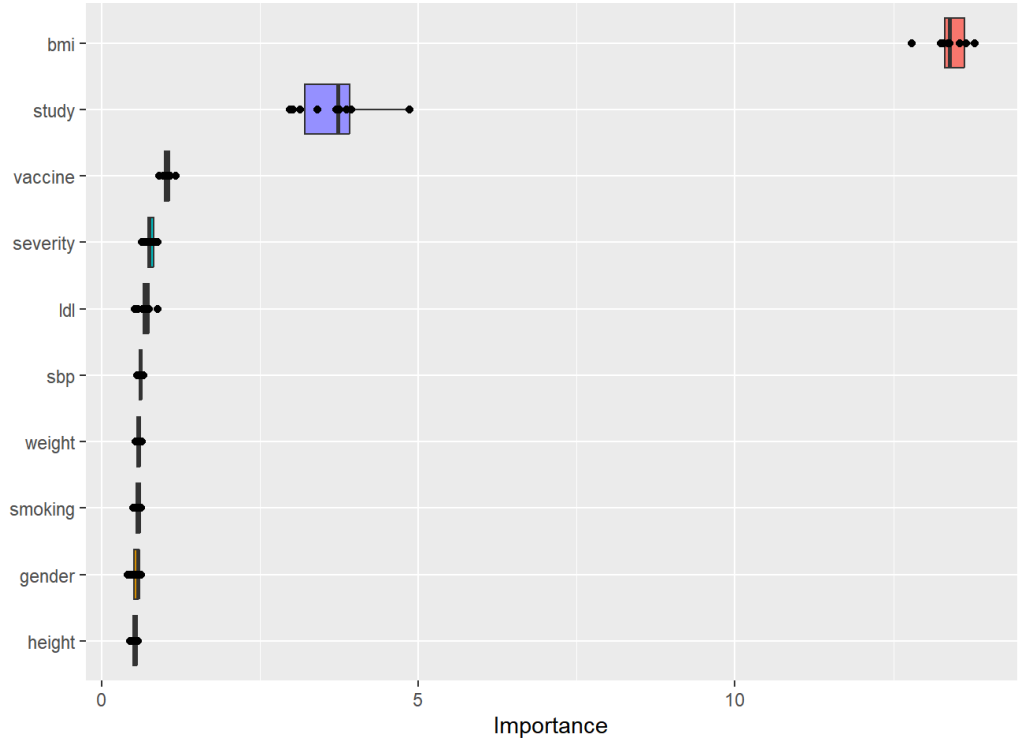

Fig.6 Variable importance plot of the Boosting classification tree