P8106 Final Report

Zhezheng Jin (zj2358)

Data Pre-processing

A total of 1000 patients with 15 variables were included in our analysis. The data was split into training data (80%) and testing data (20%). Patient ID was excluded to ensure that the random effect among sample members is negligible. Because the response variable “Severity of COVID-19 infection (severity)” in our dataset is binary, our analysis focuses on classification models that predict COVID-19 severity and understand how predictors impact the risk of severe infection. Exploratory data analysis and model training will be conducted on the training dataset. We set the seed as 2358 across the analysis.

Exploratory Data Analysis

The study includes six continuous predictors: age, height, weight, body mass index (BMI), systolic blood pressure (SBP), and low-density lipoprotein (LDL) cholesterol. Figure 1 demonstrates that the continuous predictors have a relatively symmetrical and approximately normal distribution in general. From the correlation plot in Figure 2, we can observe that there is a strong positive correlation between BMI and weight, and a negative correlation between BMI and height. SBP, LDL, and age are moderately correlated for each other. Highly correlated predictors can cause multicollinearity, potentially impacting the predictive performance of a model. To mitigate this issue, it is essential to implement further model training with cross-validation (CV).

The study includes six categorical predictors: gender, race, smoking status, hypertension status, diabetes status, vaccine status. Figure 3 displays a bar chart that shows the COVID-19 infection of 514 patients in the training dataset was not severe, while 286 patients had severe COVID-19 infection. Figure 4 demonstrates that the outcome has a similar distribution pattern among most of the predictors, which means that in general, there are more patients who had a not severe COVID-19 infection, except for those who are not vaccinated and had hypertension.

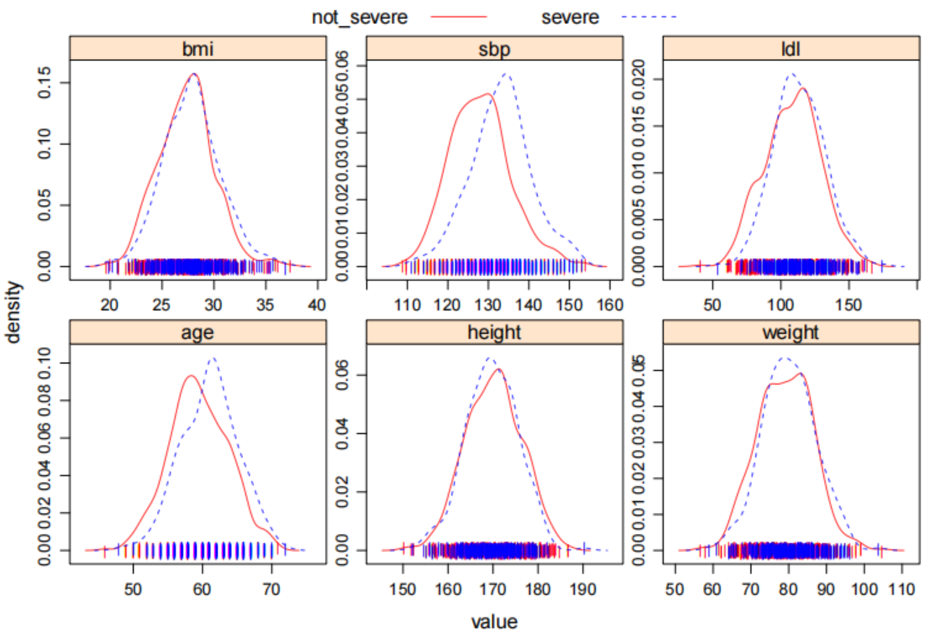

Figure 1. Density Plots of Continuous Predictors by Severity

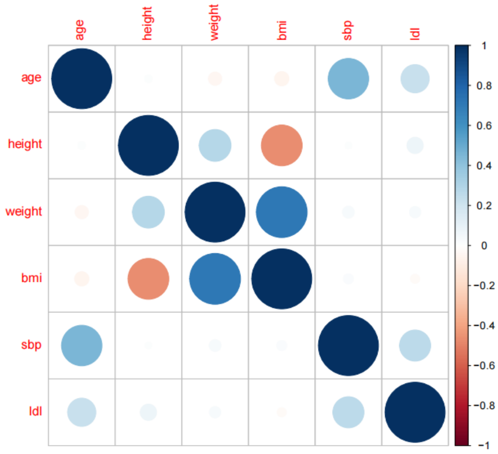

Figure 2. Correlation Plot

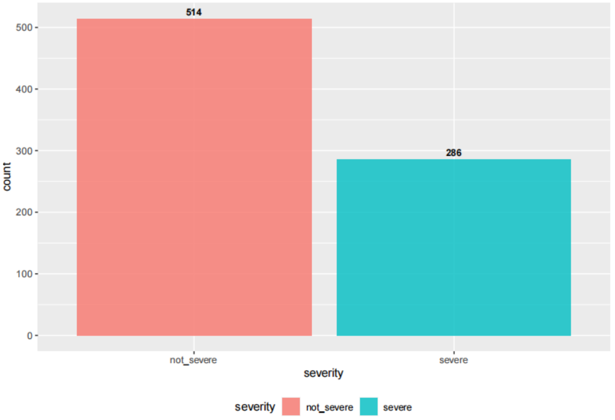

Figure 3. Bar Chart of Severity Counts

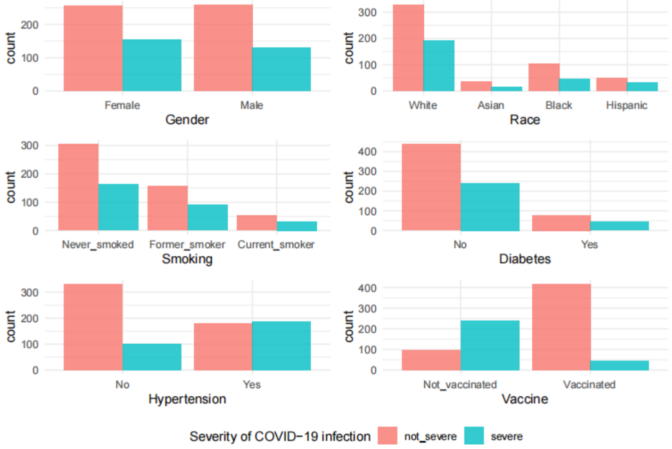

Figure 4. Bar Charts of Severity by Categorical Predictors

Model training

To identify the optimal model for predicting severity of COVID-19 infection as a binary outcome, we evaluated 5 models through model training: Penalized Logistic Regression, Multivariate Adaptive Regression Splines (MARS), Support Vector Machine (SVM) using Radial Kernel, Random Forest, and Classification Tree with AdaBoost. All models were trained using 10-fold cross validation. Following cross validation, we selected the optimal tuning parameters for each model. Additionally, we reported the test accuracy obtained by applying the final model to the test data for each model.

Penalized Logistic Regression (PLR)

The PLR model is implemented using the glmnet method in R, which specifically applies elastic net regularization. Most of the assumptions of this PLR model are the same as Linear model, except that the response variable follows a binomial distribution, where the log-odds of the outcome is a linear combination of the independent variables. After conducting cross-validation, the two optimal tuning parameters are 0.25 (alpha) from 21 candidates within range of 0-1, and 0.0324 (lambda) from 50 candidates within range of $e^{-6}\ to\ e^{3}$. The model has a test accuracy of 86%.

Multivariate Adaptive Regression Splines (MARS)

In classification, the goal of MARS is to predict the class label of each observation based on a set of predictor variables. One of the key assumptions of MARS in classification is that the classes are separable by piecewise linear functions. Additionally, MARS assumes that the class labels are mutually exclusive and that there is no overlap between classes. After conducting cross-validation, we optimized the maximum degree of interaction terms from 1 to 3 and the number of terms from 2 to 25, identifying the optimal settings as a degree of 3 and 12 terms for the best model. The test accuracy of this model is 82%.

Support Vector Machine (SVM)-Radial Kernel

SVM using Radial Kernel is an efficient algorithm for learning nonlinear functions and assumes that the data should be separable, allowing us to draw nonlinear boundaries between two classes. It also assumes independence between observations. After conducting cross-validation, we obtained the optimal tuning parameters of cost =15.41 from 50 candidates, split from e^{-2} to e^{6} , and sigma = 0.0058 from 20 candidates, split from e^{-6} to e^{-2}, which were tuned over both cost and sigma. The test accuracy of this model is 84.5%.

Random Forest (RF)

We also utilized RF for classification. RF is an ensemble learning method for classification that builds multiple decision trees during training and outputs the class that is the majority vote of the individual trees. After conducting cross-validation, we determined that the optimal tuning parameters were 2 predictors from all 13 predictors and a minimal node size of 4 from 1 to 6 under the Gini index. The test accuracy of this model is 86%.

Classification Tree(Adaboost)

AdaBoost is an algorithm that fits classification trees to weighted versions of the training data and updates the weights to better classify previously misclassified observations. It assumes independence between observations, a non-linear relationship between the outcome and predictors, roughly equal numbers of examples for each class, and that the predictors are well-normalized or standardized. After cross-validation, we determined the optimal tuning parameters to be: the number of trees is set to 3000(2000, 3000, 4000, 5000); the shrinkage parameter (λ) is set to 0.002 (0.001 - 0.003); the minimum number of observations is set to 1; and the interaction depth (d) is set to 5 (1 - 10). The test accuracy of this model is 85.5%.

Results

After training all of the models, we compared multiple models using the 10-fold CV resampling accuracy as the criterion. Figure 5 presents the comparing results of above 5 models. The results indicate that the PLR model achieved the highest median cross-validation accuracy, around 85.65%. Consequently, we selected the PLR model for the final model. The two optimal tuning parameters are 0.25 (alpha) and 0.0324 (lambda).

The final model is as following:

$$\begin{aligned}logit\left(\widehat{Severity}\right) = {} & -10.45 + 0.0437 \times Age - 0.00384 \times Height + \\ & 0.0921 \times BMI + 0.0463 \times SBP + 0.00470 \times LDL - \\ & 0.192 \times Male + 0.130 \times Current\ Smoke + \\ & 0.396 \times Hypertension - 2.69 \times Vaccinated\end{aligned}$$

In this final model, each coefficient represents the influence of a specific predictor on the likelihood of the condition being classified as severe. Positive coefficients, like age, BMI, SBP, LDL, and being a current smoker, and having hypertension, indicates that higher values of these predictors increase the probability of a severe COVID-19 infection. On the other hand, negative coefficients, such as Height and Vaccinated, suggests that taller individuals and those who are vaccinated are less likely to experience severe COVID-19 infection. The negative coefficient for male suggests that males are less likely to have severe COVID-19 infection compared to females.

After applying the trained model to the test dataset, we obtained a test accuracy of 86%. To better understand the PLR model, we created a variable importance plot. Figure 6 shows that Vaccine (vaccinated) is the most important predictor for Severity of COVID-19 infection. The rest of the predictors are all much less important than Vaccine.

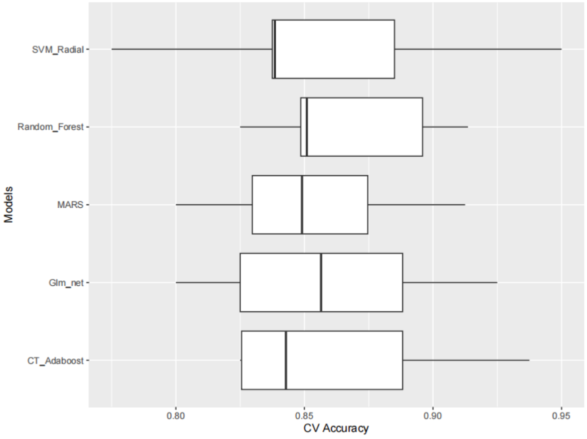

Figure 5. Cross Validation Accuracy for Training Dataset comparing 5 Models

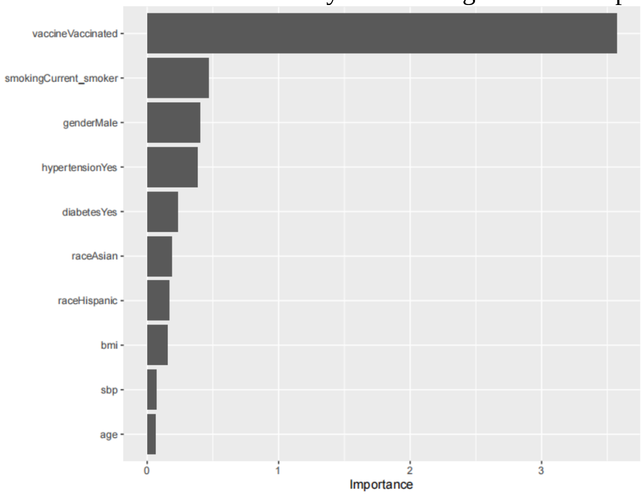

Figure 6. Variable Importance Plot of PLR model

Conclusions

In our comprehensive analysis aimed at predicting the severity of COVID-19 infection, we employed a variety of statistical modeling techniques to handle binary classification tasks effectively. Our final selection, the Penalized Logistic Regression model, demonstrated superior performance, achieving the highest median cross-validation accuracy of approximately 85.65% with a test accuracy of 86%. This model was optimized for key parameters: alpha at 0.25 and lambda at 0.0324, ensuring the best balance between model complexity and performance.

This project on predicting the severity of COVID-19 has yielded significant insights, notably highlighting the critical role of vaccination in significantly reducing the predicted severity of COVID-19 infections. Other key predictors such as age, BMI, systolic blood pressure, and hypertension were found to increase the predicted risk of severe COVID-19 infections. The Penalized Logistic Regression model, optimized and validated through rigorous cross-validation, effectively captured these relationships, providing a robust framework for assessing risk factors. These findings offer valuable guidance for healthcare strategies, emphasizing the importance of targeting high-risk groups and reinforcing the efficacy of vaccination in controlling the severity of COVID-19 infections.