What features of potential candidates should be assessed for Covid-19 vaccination?

Team Members: Siyu CHEN, Xia HU, Jia HUANG, Shuhui CHEN, Gefei LI

Submission Date: 2022.08.12

Abstract

Millions of people have been placed at risk of death since the outbreak of the COVID-19 pandemic in the United States. Researchers and clinicians have been working hard to develop the vaccine. However, it is impossible for every vaccine to be perfect or successful for everyone. People are likely to be negatively affected by the COVID-19 vaccine. The worst result is death. This project explores whether the life of a candidate is dangerous once he or she takes the COVID-19 vaccine. This project explores if variable DIED can be predicted using various features, including age, sex, vaccine manufacturer, and so on.

This project uses Logistic Regression, Decision Tree, and Random Forest in the analysis. The results suggest that the mortality rate of COVID-19 is very low. The Random Forest achieves the best predictions of the probability of death. The most important variable in the Random Forest is age.

Literature Review

With the rapid worldwide spread of COVID-19, it is of great importance to conduct an early diagnosis of COVID-19 (Zhu, 2021). The authors predict whether a patient would develop severe symptoms of COVID-19 later. In the model, they use Machine Learning techniques, such as joint classification and regression methods. The Random Forest model is used to predict the severity of COVID-19 and the possible outcome, recovery, or death (Iwendi et al., 2020). Their analysis reveals that there is a positive relationship between patients' gender and deaths. There is a high demand for Machine Learning aided systems for screening, tracking, and predicting the spread of COVID-19 and finding a cure against it (Shahid et al., 2021).

Emerging reports of deaths after the COVID-19 vaccination have raised concerns (Jain et al., 2021). A Norwegian expert group has noticed mortality in frail patients vaccinated with the Pfizer-BioNTech vaccine (Torjesen, 2021). Ingrid points out that the vaccine is likely to be responsible for the deaths of some elderly patients. It is necessary to analyze previous data and which factors affect when a vaccine is provided. The vaccination can be avoided for candidates with particular demographic factors (Sujatha, 2022).

Data

The main variables of the data exploration analysis in this project can be roughly divided into three directions(Appendix 1). The first part includes population features regarding age and sex of the sample. The second part includes DIED, which means the death count of the sample. L_threat is the indicator that measures the subjective judgment of people feeling whether they are life-threatened after vaccination. RECOVD means whether people suffering from side effects after vaccination have recovered or not. The third part is about the type of vaccine. VAX_MANU focuses on MODERNA and PFIZER since they are mainstream vaccines in the United States. VAC_DOSE_SERIES measures the total number of doses of vaccine people receive.

Some null values in this dataset have meaningful values. The DIED variable has all null values indicating people who live after vaccination Therefore, this analysis keeps the meaningful values and deletes other null variables. In addition, the variables used in this project are mostly binary features.

Approach and Methodology

First, this project uses a correlation heat map to examine the correspondences among chosen variables. Second, this project does some description to find out the characteristics of the dataset. Third, the value count of people who lived is far greater than the number of people who died in this data. This project uses SMOTE methods to address this imbalanced question.

After the data exploration, this project uses logistic regression as a baseline model and tries Decision Tree and Random Forest to find out the best model to predict. Confusion Matrix and predict score are performed to measure the level of model fitting. What is more, this project uses Hyperparameter Tuning to evaluate the model.

Results

Prediction Results

After cleaning the data and balancing samples using SMOTE, the dataset is split 8:2 into training dataset and test dataset. On this basis, use logistic regression, decision tree and random forest to train on the training set. Among them, both decision tree and random forest use five-fold cross-validation for detailed tuning adjustment parameters to avoid model overfitting. Using the optimized model to predict the test set mainly includes the following two methods：

i. Predict model category.

Using this method, the confusion matrix of the model on the test dataset can be directly obtained,and then calculate each model indicator.

ii. Predict model class probabilities.

The method is to convert the predicted category into the predicted category probability, and the obtained model data can draw the ROC curve, and then obtain the AUC value. And observe the probability of death of various groups of people.

Parameter Relationships

In order to avoid model overfitting, we perform grid search on the parameters of decision tree and random forest. For random forest, we use the following methods to perform grid search:

Table 1. Performance of Grid Search

Grid Search			
	n_estimators	Max_Depth	Max_Features

Range	start =200,
	stop=1000,num=10
		start =10,
		stop=100,num=10
			'auto','sqrt

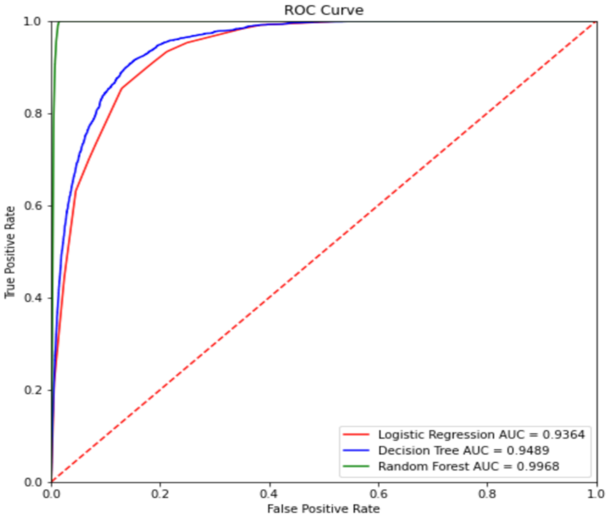
Model Evaluation Results

This project mainly uses Logistic Regression, Decision Tree, and Random Forest. Among them, Logistic Regression is the baseline model.

Table 2. Model Comparison

Approach	Precision	Recall	Accuracy	F1-Score
LR	0.861	0.906	0.882	0.883
DT	0.866	0.855	0.862	0.859
RF	0.985	0.998	0.991	0.992

Figure 1. ROC Analysis



According to the figure of the ROC Curve and results of AUC, the Logistic Regression model is the worst, and the Random Forest model is the best. Combined with the evaluation indicators, Random Forest is the best model. There may be two reasons for this phenomenon. The first is that Logistic Regression may not be suitable for this dataset. The second is that the Random Forest model may be over-fitting.

Discussion

As a result:

1) the mortality rate of this disease is very low (less than 1.5%), which is associated with age, with infants and the elderly dying at higher rates.

2) In this project, the Random Forest achieves the best predictions of the probability of death.

3) As can be seen from the important features of the Random Forest, the most important variable is age. Since the Random Forest is a tree-like model, it can be assumed that there is a relationship between age and mortality, which is probably not linear. It is consistent with the conclusions of the data exploration process.

Therefore, for policymakers, when the government is promoting the vaccination plan, they should first consider the health conditions of the elderly and children. What is better is to provide some freely required health assessments for these vulnerable groups before the vaccinations. Also, the government should continually invest in research into vaccine safety and its potential effects on humans.

Limitation

From the perspective of the model, this project adopts limited models. Therefore, more classification models can be considered to try, such as SVM and KNN. Besides, the existing models can try more parameters. The amount of data in this project is very large, but there are too many missing values to delete. A better approach is to insert variables of missing values. And more methods should be used to refine the dataset to obtain more features.

Reference

Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., Mishra, R., Pillai, S., & Jo, O. (2020). COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Frontiers in Public Health*, *8*. https://www.frontiersin.org/articles/10.3389/fpubh.2020.00357

Jain, V. K., Iyengar, K. P., & Ish, P. (2021). Elucidating causes of COVID-19 infection and related deaths after vaccination. *Diabetes & Metabolic Syndrome*, *15*(5), 102212. https://doi.org/10.1016/j.dsx.2021.102212

Shahid, O., Nasajpour, M., Pouriyeh, S., Parizi, R. M., Han, M., Valero, M., Li, F., Aledhari, M., & Sheng, Q. Z. (2021). Machine learning research towards combating COVID-19: Virus detection, spread prevention, and medical assistance. *Journal of Biomedical Informatics*, *117*, 103751. https://doi.org/10.1016/j.jbi.2021.103751

Sujatha, R. (2022). *Prediction of Suitable Candidates for COVID-19 Vaccination*. https://www.techscience.com/iasc/v32n1/45283

Torjesen, I. (2021). Covid-19: Pfizer-BioNTech vaccine is “likely” responsible for deaths of some elderly patients, Norwegian review finds. *BMJ*, *373*, n1372. https://doi.org/10.1136/bmj.n1372

Zhu, X. (2021). *Joint prediction and time estimation of COVID-19 developing severe symptoms using chest CT scan—PMC*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7547024/

Appendix

Illustration of input features and output features

<table>
<tr><td colspan="6">Input Features</td></tr>
<tr><td>Vaccine Symptoms</td><td>Description</td><td>Population Features</td><td>Description</td><td>Type of Vaccine</td><td>Description</td></tr>
<tr><td>L_THREAT</td><td>Whether people feel the life threatening</td><td>AGE_YRS</td><td>Age of samples</td><td>VAX_MANU</td><td>Vaccine manufacture (MODERNA/PFIZER)</td></tr>
<tr><td>RECOV</td><td>Whether people suffer from side effect recovered</td><td>SEX</td><td>Sex(female/ male) of samples</td><td>VAX_DOSE_SERIES</td><td>How much dosage of vaccine people accept</td></tr>
<tr><td colspan="6">Output Feature</td></tr>
<tr><td colspan="3">DIED</td><td colspan="3">Death Count (0/1), ‘0’ represents “people alive”, ‘1’ represents “people dead”</td></tr>
</table>