Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230122

Predicting death risk of COVID-19 patients leveraging machine learning algorithm

Jianming Yuan

School of Economics and Management, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

2020210964@stu.cqupt.edu.cn

Abstract. The first instance of COVID-19 was found in Wuhan, China, which mainly caused damage to human body in the form of respiratory diseases. In this study, an XGBoost prediction model was put forward according to the analysis on age, pneumonia, diabetes, and other attributes in the dataset, which was employed to estimate the COVID-19 patients' risk of death. In this study, a lot of preprocessing was carried out on the dataset, such as deleting null values in the dataset. In addition, there are strong correlation between sex, pnueumonia and death probability. In this study, XGBoost, CatBoost, logistic regression and random forest were established by machine learning method to forecast the COVID-19 patients' chance of mortality. The findings revealed that XGBoost's prediction performance was the best, while the logistic regression model performed poorly in this reported dataset of COVID-19 patients when compared to other approaches. From the feature importance map of XGBoost, it is found that age and pneumonia have great influence on the prediction of death risk.

Keywords: XGBoost, machine learning, COVID-19, pneumonia.

1. Introduction

Corona Virus Disease 2019 (COVID-19) was initially identified in Wuhan China, 2019. It is classified as a worldwide epidemic in March 2020 by World Health Organization. Vuorio, Watts et al. found that patients with familial hypercholesterolemia (FH) are more likely than healthy individuals of the same age to suffer intense COVID-19 complications [1, 2]. Li, Wang et al. furthermore discovered that age and basic illnesses including diabetes, hypertension, and others are significant risk factors for COVID-19 patients' high mortality [3]. The above two research reports show that the severity of COVID-19 patients has a great relationship with whether they have basic diseases or not. Aktar, S et al. use several machine learning algorithms, like deep learning and k nearest neighbors, and use data for several measurable clinical parameters in patients' blood samples. It is found that these parameters are of great value in forecasting the severity of COVID-19, and the accuracy and precision of their analytical methods in predicting the degree are over 90% [4]. Agrawal and Patil use chest X-ray images and patient metadata involving age and sex. Analyzed by using several different integration technologies, CatBoost is the best performing technology, Accuracy 93%, AUC 95% [5]. The former is based on the analysis of professional data, which cannot solve the problem that patients predict the risk failing to achieve the purpose of saving medical resources [6, 7]. In this work, the fundamental physical state of COVID-19 patients and the XGBoost prediction model were used to estimate the risk

DOI: 10.54254/2755-2721/8/20230122

of death for COVID-19 patients. Boosting is an ensemble technique that adds additional models to existing models to fix flaws in the original models. Models are gradually added till no more advancements are possible. XGBoost performs very well for classification or regression problems. As a kind of boosting models, it is widely used in industrial fields, which could be attributed to its outstanding performance, simplicity, and limited execution time. It is improved on the basis of the original GBDT. It automatically uses the multi-threading of the CPU to build trees in parallel, and at the same time improves the algorithm to improve the accuracy. Compared with CatBoost, logistic regression, random forest has better model efficiency and calculation speed. In this work, the effectiveness of the XGBoost is compared with other famous classification algorithms for the death risk estimation of COVID-19. It is a fair comparison which could on one hand demonstrate the effectiveness of the XGBoost model and on the other hand show how the machine learning algorithms performs on the specific COVID-19-related problem.

2. Method

2.1. Dataset and processing

The dataset of this study comes from kaggle website and is provided by Mexican government [8]. It contains 1,048,576 COVID-19 patient data, with a total of 21 characteristics, among which the data type of data_died is date, age is discrete data, and other characteristics are classified data. Data_died is replaced by a new feature death, and then data_died, age over 110 years old and abnormal values are deleted. There are some null values in the data. Because of the big data and the large amount of null data, null values of related features are processed to reduce the influence of null values on prediction results and accuracy. Before data processing, patient_type was used to replace death risk, the absolute value of correlation coefficient between other attributes and this attribute was observed, and it was found that the absolute value from high to low was as follows: intubed＞ICU＞age＞medical_unit＞classification_final. After data processing, death was used to replace death risk, and the correlation coefficient between other attributes and this attribute was observed, and it was found that the absolute value from high to low was as follows: patient_type＞pneumonia＞age＞diabetes＞hypertension＞classification_final.

The algorithm could automatically learn from training data, and the parameters of fitting curve were determined to achieve the purpose of establishing the model. To measure the performances of the implemented models, a series of indexes are calculated on the test data. Eighteen percent of the patients are clustered for training, while the remaining twenty percents of the patients are utilized as the test set because there were no local abnormalities in the dataset due to its enormous capacity.

2.2. Feature screening

By limiting the number of features, feature selection could simplify the model and increase its ability to generalize [9]. Since there are still 17 features in the dataset after data processing, it is very important to obtain effective features. This study is based on the XGBoost algorithm [10], which can be directly through the decision tree method for feature screening.

2.3. XGBoost

XGBoost is a representative boosting model, which iteratively produces weak classifiers through multiple rounds. During the training processes, each classifier is gradually improved leveraging the residual information produced from the latest trained classifier. It is required that weak models should possess low variance and high bias. Weak classifiers generally choose CART TREE (XGBoost also supports linear classifiers). Each classification regression tree won't have a particularly deep root system. To determine the final total classifier, the weak classifiers from each training round are weighted summarized. After multiple iterations and second-order Taylor expansion of XGBoost, the objective function is formula (1).

DOI: 10.54254/2755-2721/8/20230122

$$obj^t = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega\left(f_t\right) + constant \quad (1)$$

2.4. Experiment

The early XGBoost prediction model was constructed and trained by using the training data. Parameters of XGBoost were adjusted and optimized by grid search. In this study, the accuracy, precision, F1 value, recall rate, and a threshold free index called area under curve for assessing the built model's prediction. At the same time, it was compared to the machine learning algorithms CatBoost, logistic regression, and random forest.

3. Result

The effectiveness of XGBoost model is validated in this part, for demonstrating its effectiveness on the death risk prediction problem. Afterwards, it is compared to other classification models to further illustrate its superiority.

3.1. Results of the prediction model

Extreme gradient boosting, sometimes referred to as extreme gradient lifting tree, is a boosting method implementation, which mainly reduces the error of the model and finally reduces the gap between the model's actual value and its predicted value. In this paper the effectiveness of it is validated on the death risk prediction problem. From Figure 1 and Table 1, when there are 17 eigenvalues, the accuracy value is the highest. The corresponding accuracy is 0.9429, F1 is 0.9695, precision is 0.9607, recall is 0.9784 and AUC is 0.733.

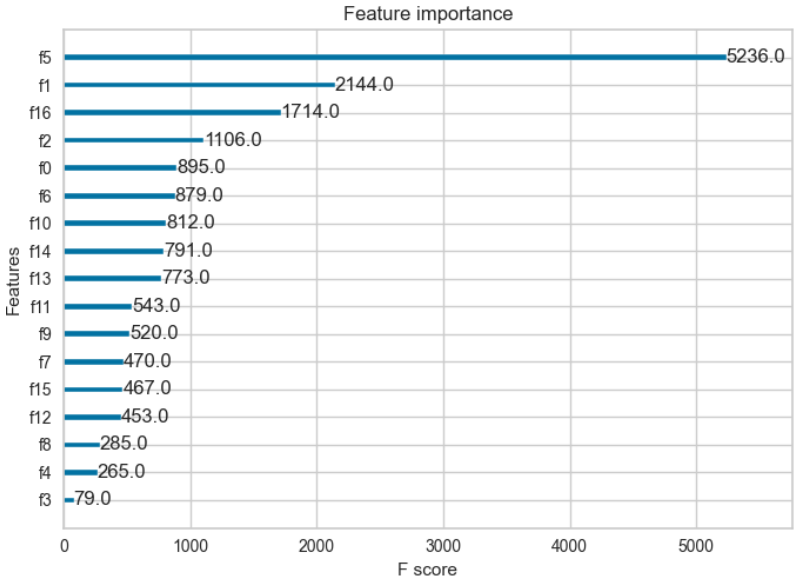

Figure 1. Feature importance.

Table 1. Feature number and Accuracy value.

Thresh	0	0	0	0.001	0.001	0.001	0.001
n	17	16	15	14	13	12	11
Accuracy	94.29%	94.25%	94.28%	94.26%	94.28%	94.28%	94.28%
Thresh	0.001	0.001	0.001	0.001	0.002	0.004	
n	10	9	8	7	6	5	
Accuracy	94.27%	94.25%	94.23%	94.18%	94.18%	94.17%	

DOI: 10.54254/2755-2721/8/20230122

3.2. Evaluation of the XGBoost

XGBoost trains and tests the selected optimal feature subset, and the prediction result of the model was better. Accuracy was 0.9429, F1 was 0.9695, precision was 0.9607, recall was 0.9784 and AUC was 0.733. For R, the same conclusion can be drawn from Figure 2.

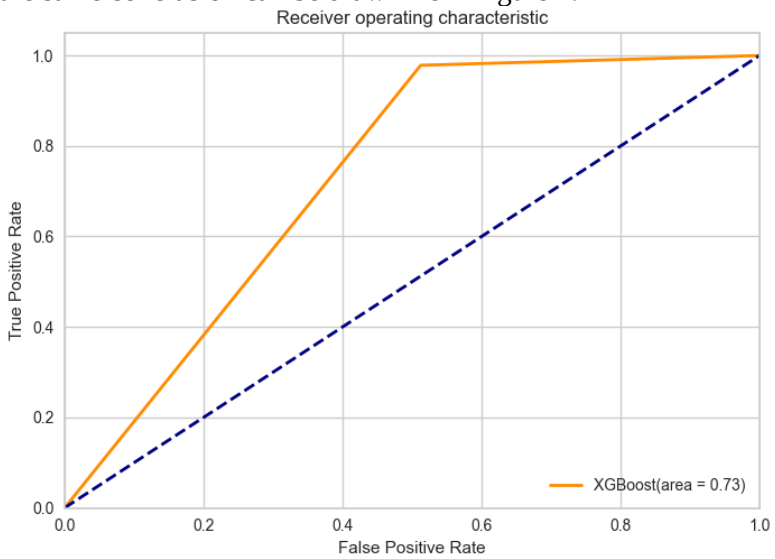

Figure 2. ROC curve of the XGBoost result.

3.3. Comparison of other machine learning models

Test sets were used to compare the performance of XGBoost, CatBoost, Logistic Regression and Random Forest. The values of different measurements of the above four models are shown in Table 2. XGBoost surpasses other models and achieves the performance, with an accuracy of 0.9429, an F1 of 0.9695, a precision of 0.9607, a recall of 0.9784, and an AUC of 0.733. With an accuracy of 0.9380, an F1 of 0.9669, a precision of 0.9573, a recall of 0.9768, and an AUC of 0.7089, the logistic regression model has the worst prediction performance. Meanwhile, it has the slowest running speed compared to other models.

Table 2. Performance indicators of each model.

	accuracy	f1	precision	recall	AUC
XGBoost	0.9429	0.9695	0.9607	0.9784	0.733
CatBoost	0.9427	0.9694	0.9595	0.9796	0.7247
RandomForest	0.9339	0.9645	0.9597	0.9694	0.724
LogisticRegression	0.9380	0.9669	0.9573	0.9768	0.7089

4. Conclusion

Although the successful development of vaccines and specific drugs will reduce the risk of death of patients to a certain extent, simpler and more accurate prediction models are needed to help patients reduce the risk of death. In this study, the XGBoost prediction model I got by training the basic disease information of the patient's body is the best. At the same time, from the feature importance map, it is found that age and pneumonia have great influence on the prediction results of the model. Older patients with pneumonia should avoid being infected. If infected, they should go to the hospital immediately. Other patients with low death risk can be cured by their own immune system to avoid excessive occupation of medical resources. Finally, the goal of predicting death risk of COVID-19 patients through their own basic disease information was achieved. Besides XGBoost, neural network, as a popular classifier, has achieved good results in many other tasks. However, it consumes a lot of

DOI: 10.54254/2755-2721/8/20230122

computing resources, so it is not used in this paper. In the future, if there is a stronger experimental platform, the performance of other models such as neural networks will also be verified.

References

[1] Vuorio, A., Watts, G. F., & Kovanen, P. T. (2020). Familial hypercholesterolaemia and COVID-19: triggering of increased sustained cardiovascular risk. J Intern Med, 287(6), 746-747.

[2] Baloch, S., Baloch, M. A., Zheng, T., & Pei, X. (2020). The coronavirus disease 2019 (COVID-19) pandemic. The Tohoku journal of experimental medicine, 250(4), 271-278.

[3] Li, X., Wang, L., Yan, S., Yang, F., Xiang, L., et, al. (2020). Clinical characteristics of 25 death cases with COVID-19: a retrospective review of medical records in a single medical center, Wuhan, China. International Journal of Infectious Diseases, 94, 128-132.

[4] Aktar, S., Ahamad, M. M., Rashed-Al-Mahfuz, M., Azad, A. K. M., Uddin, S., et al. (2021). Machine learning approach to predicting COVID-19 disease severity based on clinical blood test data: statistical analysis and model development. JMIR medical informatics, 9(4), e25884.

[5] Bharathi, M. L., Aravindan, J., Venusamy, K., & Basha, R. F. K. (2022). Artificial Neural Network based Automatic Prediction Unit for COVID 19 in Asymptomatic Patient. In 2022 International Conference on Advances in Computing, Communication and Applied Informatics, 1-5.

[6] Esai Selvan, M. (2020). Risk factors for death from COVID-19. Nature Reviews Immunology, 20(7), 407-407.

[7] Fridman, S., Bullrich, M. B., Jimenez-Ruiz, A., Costantini, P., Shah, P., et, al. (2020). Stroke risk, phenotypes, and death in COVID-19: systematic review and newly reported cases. Neurology, 95(24), e3373-e3385.

[8] Eduardo, R., (2020), COVID-19 in Mexico, URL: https://www.kaggle.com/code/lalish99/covid-19-in-mexico

[9] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM computing surveys (CSUR), 50(6), 1-45.

[10] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., et, al. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.