2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE 2021)

2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) | 978-1-6654-1540-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICBAIE52039.2021.9389914

Autism Classification Based On Logistic Regression Model

Yuanrui Zheng[1]
Yunnan Normal University Business School,
Yunnan, China,
aaronzheng87@gmail.com

Tingyan Deng[1]
Vanderbilt University,
Nashville, Tennessee,
tingyan.deng@vanderbilt.edu.

Yaozheng Wang*
Victoria University
Henan, China,
yaozheng.wang@live.vu.edu.au

***Abstract*— Autistic Spectrum Disorder (ASD) is a developmental disability can affect communication and behavior. Existing research has shown that early diagnosis can help doctors to find this disease early and can save significant healthcare costs. With the rapid growth of ASD cases, a ASD related dataset created for scientists and doctors to investigate this disease. Autistic Spectrum Disorder Screening Data for Adult is a well-known dataset, which contains 20 features to be utilized for further analysis. This article developed and test an Autism classification algorithm which based on logistic regression model. The result of this study provided a model can predict the ADS in an average F1 score of 0.97, which displays the superiority of proposed model. Besides, the data visualization part displays several feature distribution images for people to better understand the data and related feature engineering.**

***Index Terms*—ASD, Logistic Regression, Classification**

I. Introduction

Autistic Spectrum Disorder (ASD) is a mental disorder can lead to barriers of society, affect cognitive and communication. Existing researches were investigated to recognize it, prevent it or even treat it. In work [1], the authors presented a method using the screening trying to diagnosis this disorder. Their work has huge impact and has cited by more than 1600 times. In work [2], LT Curtis and his partners gave several approaches including nutritional and environmental approaches to prevent the Autism. In work [3], music therapy are used to enable communication and expression and thus can solve some problems of this disorder.

A. Related Work

In recent years, there have been an acceleration of concern in recognizing ASD by methods, which are based on machine learning. Generally, the models to solve this problem are all classfication ones. In paper [4], F Thabtah give a summary of recent machine learning methods in autistic spectrum disorder. In [5], F Thabtah considers the ASD diagnosis problem as a classification problem. He assumes the model are plugged into a screening tool to accomplish one or more of the aforementioned goals. In [6], Hyde K shows applications of supervised machine learning methods. In this paper, we present an Autism classification based on logistic regression model.

Logistic regression model is a classification model, which uses a logistic function to model a binary dependent variable [7]. Logistic function is the inverse of the natural logit function, which can convert logarithm of odds into probability. Logistic regression algorithms as well as other classification algorithms like support vector machine [8] and random- forest [9]. These algorithms have the ability to fit the training data and inference from test data. In this scenario, we utilize the logistic regression as the fitter to learn the mapping relationship between input data, screen data, to classification results.

B. Our Contribution

Our work makes a practical contribution about applying logistic regression to ADS diagnosis. More specifically, the data columns and meanings are described in the first step. Then the data imputation method or feature engineering methods are introduced to further proceed the original data. Besides, the data visualization step contains many figures are shown for a better understanding. Finally, we briefly explain the logistic regression model and how to fit the model with processed data. The experiments show the metrics like accuracy, recall, and F1 score. We got an average F1 score of 0.97, which proves the feasibility of our model.

The remainder of this paper is organized as follows. Section II introduces data structure and feature engineering methods. Besides, several distribution figures of different features are shown in this section. Section III explains logistic regression model, and Section IV displays experimental results and analyzes. Finally, Section V gives the summary of whole paper.

II. Data And feature engineering

The data set can provide a lot of information. In order to explain the dataset intuitively, we list the features of dataset in the table 1.

TABLE I. Data Structure

A1_Score to A10_Score
age
gender

978-1-6654-1540-8/21/$31.00 ©2021 IEEE

ethnicity
jundice
austim
contry_of_res
Used_app_before
result
relation
Class/ASD

The features age, gender, ethnicity are attributes of the ASD testers and easy to figure out the meaning of them. The A1_Score to A10 are the answer code of the question based on the screening method used.

For saving memory cost, the data type should be adjusted. The floating-point numbers in original data set take up a lot of memory. In this paper, these data are compressed into a smaller floating-point number type.

Time period has a great influence on sales fore-casting, so the more detailed the time feature extraction, the better. A lot of information about time can be extracted from the timestamp. For example, through processing, it can be determined that which week, month, quarter a certain day is in, and whether it is a weekend. You can also get the lunar date of a certain day through conversion.

Two types of statistical characteristics are calculated in program, 1) rolling characteristics, statistical characteristics within a certain period; 2) lag characterrristics, characteristics ending at a certain time point, such as the characteristics of the previous 7 days, the previous 14 days, and the previous 21 days. For commodity prices, statistics such as maxi-mum, minimum, and median value are calculated

However, these feature columns have some missing values, Nan values. In the Figure 1, the missing values are noted are yellow in the figure.

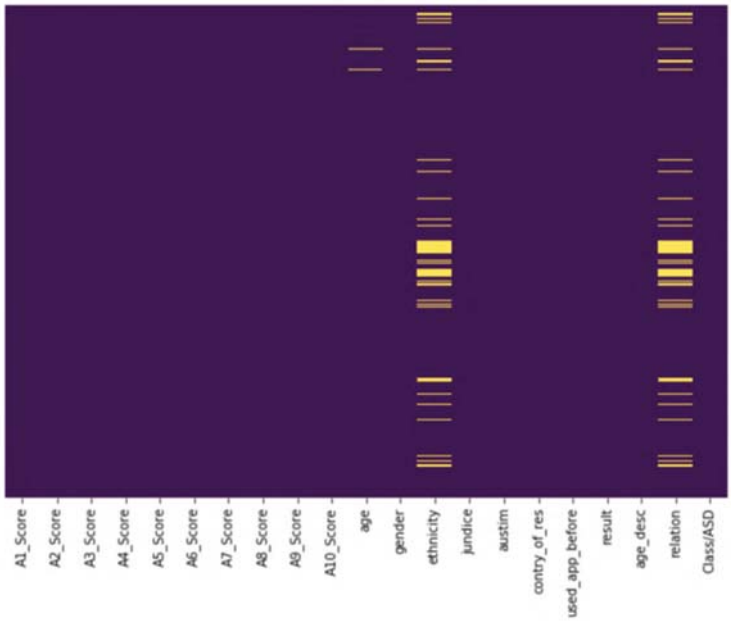

Figure 1. The feature columns

To better understand the data and related feature engineering, we display several feature distribution images. And the age distribution of the ASD patient in Autism Screening Classification data set is shown in Figure 2. Since the average age of the testers is around 30, we use 30 to represent the testers of whom the age information is absent. And the age distribution shown in Figure 2 is fixed. Macroscopically, the proportion of testers decreased with age. Specifically, the 20 to 30 years old account for more testers than 30 to 40, 40 to 50, and 50 to 60.

The Proportion of male and female patients with ASD is shown in Figure 3. According to Figure 3, we can notice that there are more female testers who suffer from Autistic Spectrum Disorder than male testers. And there are more male testers who are not ASD patients than female testers. After simple math computation, thus, the proportion of ASD patients in female testers is large than that in male testers. It indicates that gender information might be an important feature in ASD diagnose.

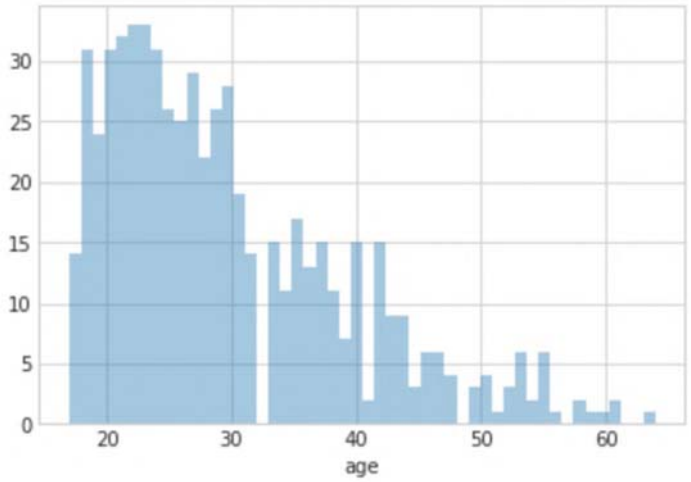

Figure 2. The age distribution of the ASD patient in Autism Screening Classification data set.

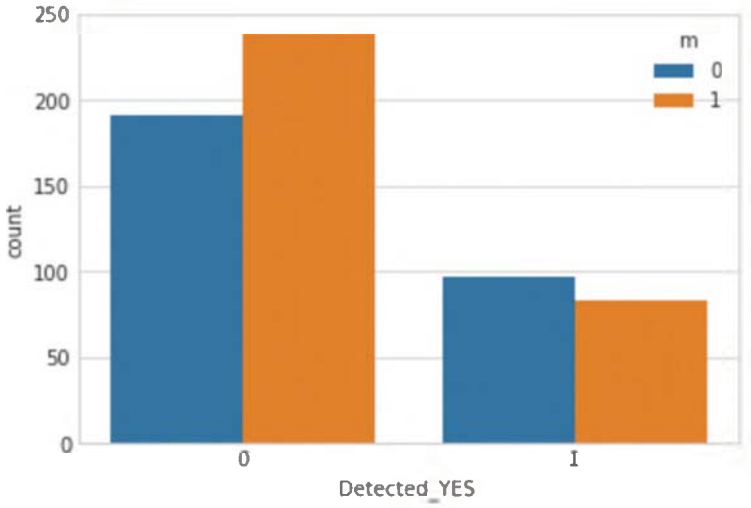

Figure 3. The Proportion of male and female patients with ASD.

III. LOGISTIC REGRESSION MODEL

In this section, the logistic regression model is briefly presented. And the method to fit the model with the processed data is described as well.

Logistic regression model is a type of classification model. The key idea of the model can be expressed as

$$f(z)=\frac{1}{1+e^{-z}}=\frac{1}{1+e^{-\boldsymbol{\theta}^T\mathbf{x}}},$$

Where $\boldsymbol{\theta}$ indicates the model parameter vector which we want to get. And $\mathbf{x}$ indicates the input variable vector. The input variable vector contains the factors we assumed to affect the result $f(z)$, which is the possibility of ASD diagnosed. The model parameter vector is estimated during training, which is called model fitting as well. During training, a number of input variable vectors and there corresponding results which are usually annotated manually are used to estimate the model parameter vector. And maximum likelihood estimation is used in the training process. There are several optimization methods which can be used in model parameter estimation. These methods are liblinear, LBFGS, Newton-CG, and SAG. In our experiments, the liblinear is used, which utilizes coordinate descent method to conduct optimization. Besides, we utilize a L2 penalty to avoid over fitting.

After proper training, the model with the model parameters can be an estimator for our application. And during testing or application, we can feed the model with unlabeled variable vectors and the model will output the corresponding results it estimates.

It can be regarded as a combination of linear regression and a sigmoid function. Although the concept of it looks simple, the proper application of it can solve a few complicated problems.

IV. Experimental results and analysis

The experimental results and the analysis of it are present in this section.

To quantitatively evaluate the results of our work, we utilize precision, recall rate, and F1 score as the metrics. The precision is expressed as

$$\text{precision} = \frac{TP}{TP + FP}$$

Where TP indicates the number of real positive samples, and FP indicates the number of false positive samples. Meanwhile, the recall rate is expressed as

$$recall = \frac{TP}{TP + FN}$$

Where FN indicates the number of false negative samples. And the F1 score is expressed as

$$F1 = \frac{precision \times recall}{precision + recall}$$

Which is a comprehensive measure of precision and recall rate.

By conducting experiments on Autism Screening Classification data set, we can obtain results shown in Table 2.

As the table shown, the logistic regression based method has obtained 0.97 precision in the experiments. Which indicates the testers who are predicted to suffer from ASD by the method are possibly ASD patients.

Meanwhile, the logistic regression based method has obtained as high average recall rate as 0.97. It means that the method can possibly predict an ASD patient if the tester is suffering from ASD.

The average F1 score of the method is high as well. It reveals the efficiency and robustness of the logistic regression method. The method maintains a good balance between precision and recall rate and approaches high levels on both.

The high performance of the method on the data set illustrates the data and feature engineering work well. And the logistic regression model has learned accurate model parameters. It shows the efficiency of the training process.

TABLE II. Quantitative results of the experiments on Autism Screening Classification data set.

	precision	recall rate	F1 score
1	0.98	0.97	0.98
0	0.92	0.96	0.94
average	0.97	0.97	0.97

V. Conclusions

As a a developmental disability, Autistic Spectrum Disorder has an influence on the communication and behavior of the patients. Early detection can highly reduce the incidence of the disease. In this work, we apply the logistic regression model in the ASD diagnose process. This application includes data and feature engineering, model training and model testing. After proper work on these parts, our work has achieved high performances on the Autism Screening Classification data set, which is indicated by the high values on precision, recall rate and F1 score. It reveals that this machine learning based method has potential to help ASD diagnose in practice. Furthermore, we exhibit several feature distribution images to help the understanding of the data and feature engineering in the visualization part.

References

[1] Filipek P A, Accardo P J, Baranek G T, et al. The screening and diagnosis of autistic spectrum disorders[J]. Journal of autism and developmental disorders, 1999, 29(6): 439-484.

[2] Curtis L T, Patel K. Nutritional and environmental approaches to preventing and treating autism and attention deficit hyperactivity disorder (ADHD): a review[J]. The Journal of Alternative and Complementary Medicine, 2008, 14(1): 79-85.

[3] Gold C, Wigram T, Elefant C. Music therapy for autistic spectrum disorder[J]. Cochrane Database of Systematic Reviews, 2006 (2).

[4] Thabtah F. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward[J]. Informatics for Health and Social Care, 2019, 44(3): 278-297.

[5] Thabtah F. Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment[C]//Proceedings of the 1st International Conference on Medical and health Informatics 2017. 2017: 1-6.

[6] Hyde K K, Novack M N, LaHaye N, et al. Applications of supervised machine learning in autism spectrum disorder research: a review[J]. Review Journal of Autism and Developmental Disorders, 2019, 6(2): 128-146.

[7] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. Vol. 398. John Wiley & Sons, 2013.

[8] Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." Neural processing letters 9.3 (1999): 293-300. Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001: 1189-1232.

[9] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.