Comparison of Logistic Regression and Decision Tree Models for Mental Health Estimation of Employees

Muyun Li[1]

[1]College of Arts and Science, Vanderbilt University, Nashville, Tennessee State, 37235, USA
muyun.li@vanderbilt.edu

Abstract. Mental health accompanies every human being inevitably and has great significance in helping people address life stress and realize their abilities. However, mental health is also a double-edged sword, which mental health issues can hinder people from carrying out daily activities normally and keeping in a good mood. Not only should the general public be aware of the importance of their mental health, but also those industries that rely on human resources should pay special attention to their employees' mental health in order for the normal operation of the essential tasks. This paper aims at constructing feasible models helpful for normal people to predict their own mental state and organizations to predict their employees' mental health state. To predict the mental health state, this paper examines two models of logistic regression and decision tree classifiers. The results indicate that logistic regression is relatively stable but not perfect in accuracy, positive predictive value, and true positive rate while decision tree classifiers are excellent at positive predictive value but poor at true positive rate.

Keywords: Mental Health, Logistic Regression, Decision Tree.

1 Introduction

World Health Organization explains mental health to be "a condition of well-being in which human beings recognize their own capabilities, are able to handle the typical stresses of living, operate effectively and efficiently, and are capable to contribute to making a difference in the lives of others." [1]. From the above perspective, it could be noticed that mental health accompanies every human being inevitably, and it has the ability to affect every individual positively or negatively in a variety of aspects of his or her daily life. For instance, research has shown that in most high-income nations, mental illness is now the main reason for illness-related absences and disability benefits, and the increasing costs to society and the economy make health and employment a higher policymaker priority [2]. Furthermore, mental health is prone to change with transitions between different stages in people's lives, like children to teenagers and teenagers to adults. For example, university enrollment may be a cause of anxiety and an acute trigger of stress. Academic requirements increase, and new social relationships are formed. Moving from high school to university reduces contact and, likely, support from close companions and relatives for students

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_3

who transfer away from home. Difficulties in managing these transition-related stressors may result in diminished academic performance and higher levels of psychological distress [3].

Recognizing its nearly ubiquitous significance and influence, it is consequently essential to employ means to measure and predict mental health, especially for corporations or other human-related industries that depend significantly on human resources for their everyday production and functioning. It is especially noticeable that in this contemporary period, right after or still throughout the COVID-19 epidemic, the general population's mental condition is overall unstable since when the environment alters, people typically feel apprehensive and insecure [4]. COVID-19 has increased the prevalence of identified risk factors for mental health issues like social estrangement, loneliness, idleness, and greater availability of alcohol as well as virtual wagering resulting from quarantine along with physical separation [5]. Moreover, researchers have shown that in addition to post-traumatic stress disorder associated with recovery from a potentially fatal physical illness, it appears that stigma, monetary damages, and employment insecurity may have a long-term impact following COVID-19, which may induce more anxiety and mental distress on employees and affect their working efficiency [6].

To better alleviate the loss brought by adverse mental health conditions, this work employs the means of machine learning that implements complex mathematical algorithms that enable the analysis of multiple variables and their relationships, in addition to the evaluation of the accuracy of a mathematical framework [7]. Additionally, machine learning has other advantages like lower cost and greater efficiency in comparison to mental health counselors, which make it more suitable for companies with a large-scale of employees. This article focuses on adopting machine learning algorithms to predict individuals' mental health conditions based on a wide range of data, which is more objective and aims to evaluate the subjects' mental health conditions better.

2 Method

2.1 Dataset

The data set is from a survey conducted in 2014. The survey collected the respondents' age, gender, country, prior employment status, family history of mental illness, mental health history, etc., and it asked the respondents' willingness to seek treatment for a mental health condition and if they feel that mental health condition interferes with their work. Based on the data collected, at first glance, there is a total of 1259 sample respondents and 60% were residents of the US, while 15% were residents of Britain. After filtering out invalid data, like those respondents with ages beyond common sense or under the legal age for employment, the rest of the data used for later machine learning contained respondents with age ranges from 18 to 72.

2.2 Models

The two machine learning methods used to generate and compare the results are logistic regression and decision tree classifier. Logistic regression can be advantageous in probability forecast given that (compared to log-binomial regression, for example) the model is mathematically limited to generating probabilities on an interval [0,1] while typically merges on parameter estimates with relative ease. The advantage of logistic regression is that it is simpler to manage multiple explanatory variables all at once, and one may employ continuous explanatory variables. Additionally, logistic regression makes it possible to eliminate confounding effects by examining the relationship between all variables simultaneously [8]. Another significant advantage of logistic regression is that in examining binary outcomes, logistic regression preserves many characteristics of linear regression [9]. For the method of decision tree classifiers, they serve as methods of categorization that define a "tree" of cut points that optimize a certain amount of variation among the final nodes of the tree. All other nodes then represent comparatively uniform classes. For their advantages, they are thoroughly researched and simple for understanding as well as utilize, and decision tree program is readily accessible in typical programs [10]. Decision trees are rigorously non-parametric and do not necessitate assertions about the input data distributions. Furthermore, decision trees are capable of managing complex interactions among characteristics and categories, data absence, as well as quantitative and classified inputs in a natural manner. Given that the classification structure is clear and consequently readily comprehensible, decision trees have a significant natural appeal [11]. To compare these two methods to see which one provides greater prediction accuracy, this paper visualizes the results through a confusion matrix, which a classification system's confusion matrix includes data about real and forecasted classifications [12].

3 Result

The four confusion matrices, demonstrated in Fig 1, are the output results of altering one parameter of the logistic regression models and extra tree classifiers with all else equal. The first column represents the predicted positive (PP) values, and the second column represents the predicted negative (PN) values. The first row represents the actual positive (P) values, and the second row represents the actual negative (N) values. The two rows represent predictive positive and negative values, and the two columns represent actual positive and negative values accordingly. The cell of the first row and the first column is the true positive (TP) value, which is the correct result since the predicted value corresponds to the actual value and is both positive. The cell located in the first row along with the second column is a false negative (FN) error, which is the incorrect test result indicating that a condition does not apply and is also called a type II error. The cell of the second row and the first column is the false positive (FP) error, which is a result that indicates a given condition exists when it does not and is also called a type I error. The cell at the intersection of the second row and second column is called the true negative (TN), which is another correct

result in which the predicted value corresponds to the actual value and is both negative.

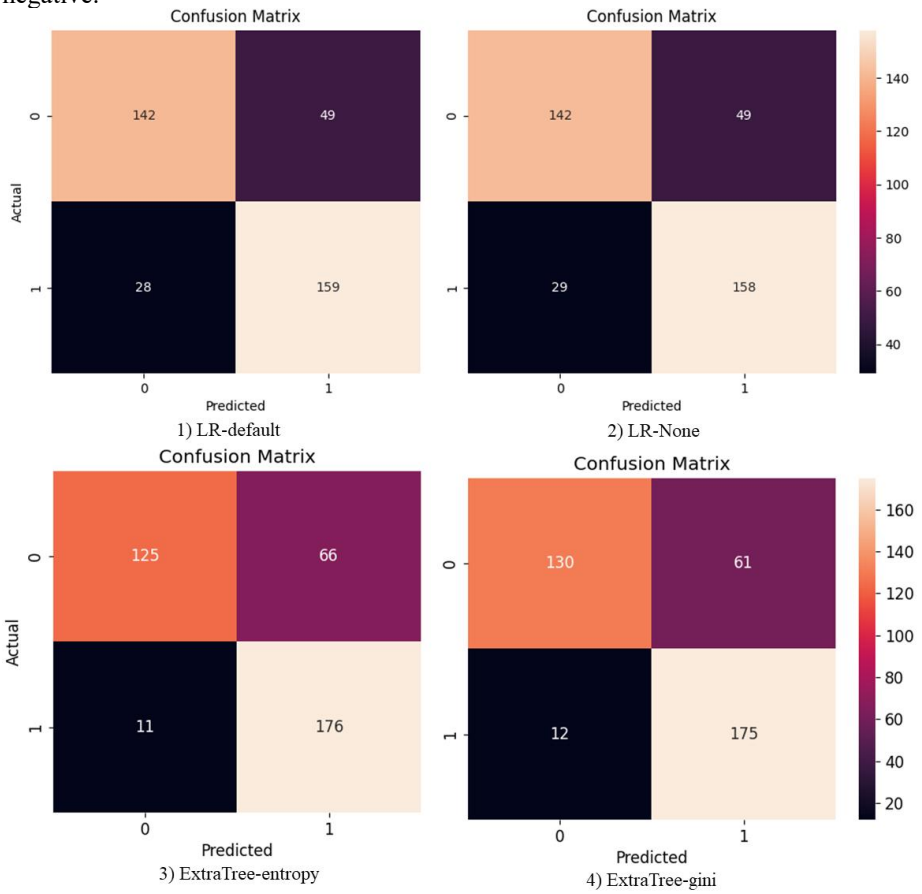

Fig. 1. Confusion matrixes of compared models (Picture credit: Original).

In the comparison results displayed in Table 1, accuracy (ACC) is defined as the ratio of exact predictions (true positives and true negatives) within all examined cases. The greater the accuracy, the better the prediction ability the model has. Seen from the calculation results for the four models, it is evident that the decision tree classifier with the criterion "gini" parameter generates the most incredible accuracy. However, the accuracy is nearly identical, with only tiny variations for the four models.

Table 1. Result comparison of different models.

	LR-default	LT-none	ExtraTree-entropy	ExtraTree-gini
Acc	0.7963	0.7937	0.7963	0.8069
Precision	0.8353	0.8304	0.9191	0.9155
Recall	0.7335	0.7435	0.6545	0.6806

Positive predictive value (PPV), also called precision, is defined as the proportion of true positive results in statistics and diagnostic tests. A high outcome may be understood as an indication of the reliability of the statistic in question. From the chart, it could be observed that decision tree classifiers generate higher positive predictive value than logistic regression, which means that it could be concluded that decision tree classifiers are more accurate in predicting the positive values.

True positive rate (TPR), likewise known as sensitivity, refers to the likelihood of obtaining a positive outcome from a test assuming the subject is, in fact, positive. In this case, sensitivity can be interpreted as the test's ability to correctly detect respondents with mental health issues out of those who display the cues or possible symptoms. From the data in the chart, it is evident that the logistic regression model performs better than decision tree classifiers, which means that given the condition that the respondents have mental health issues, logistic regression has more outstanding performance or accuracy in correctly detecting these people than decision tree classifiers do.

4 Discussion

It is worth discussing why altering certain parameters of the two models would result in different confusion matrices, accuracy, positive predictive value, and true positive rate. For the logistic regression model, the parameter of “penalty” refers to penalized logistic regression, which penalizes the logistic model for having excessive variables. This causes the coefficients of the fewer significant variables to approach zero [13]. The default setting of “penalty” is “l2”, which is also called a “ridge regression” and intends to eliminate possible problems of over-fitting by bringing the coefficient’s “squared magnitude” as the penalty term to the loss function. The other setting this paper includes of “penalty” is “none,” meaning that no penalty term is added to the loss function. From the result generated by altering the “penalty” parameter, it is apparent that adding the “penalty” makes little difference in improving or worsening the final result. In other words, this model performs nearly equally decent in its predictions with different parameters, though the accuracy of 0.79 is still not high enough to be considered accurate.

For the decision tree classifiers, this paper compares the result by altering its parameter of “criterion” from “entropy” to “gini.” Specifically, the parameter “criterion” decides how the impurity of a split or a node will be assessed. The “gini” criterion determines the likelihood that a randomly selected instance will be classified incorrectly. In contrast, the “entropy” criterion evaluates the level of disorder in a node, and a node with more variable composition has higher entropy. The results generated by different parameters of the decision tree classifiers are also similar, and it is noticeable that this model results in an exceptionally high positive predictive value but a relatively low true positive rate. Consequently, though not the perfect model, the decision tree classifiers can be adopted to predict whether respondents have mental health issues.

This paper identifies areas for future research to enhance the model's accuracy, like adopting a more comprehensive range of the data and more updated data can be used since the data this paper is based on are before the COVID-19 pandemic period,

which failed to take into account of the possible influence of COVID-19 on the respondents' answers. Additionally, future surveys are recommended to target a more specific population (like people in a certain state or a particular country) instead of a random group of cyber citizens so that a more accurate model specific to certain groups of people can be constructed with greater reliability. Furthermore, survey questions with the form of limited choices instead of short responses are recommended (like limiting the gender choices to male and female and age choices to several ranges instead of manually input), which can effectively eliminate the invalid answers and makes it more convenient to categorize data collected. Moreover, other questions that are more related to respondents' mental health state can be added to the survey so that the researchers can evaluate if the answers from indirect measures (like if the respondents would seek help if they have mental health issues) correspond with those from direct measures (like self-evaluation of past or current mental health state), which allows further analysis and improvement of the model used to predict the respondents' mental health.

5 Conclusion

In conclusion, given that mental health plays a vital role in many aspects of people's daily lives, it is essential for everyone to keep a good mental state and find the problem in time if there is a potential sign for mental health issues. In order to predict if people have mental health issues, this paper adopts the models of logistic regression and decision tree classifiers. The results indicate that neither of the two models is perfect in correctly detecting respondents with mental health issues out of those who display the cues or possible symptoms and accuracy with high enough exact predictions within all examined cases. Additionally, it is hard to determine which model is comparatively better since the logistic regression performs similarly in all three indicators. At the same time, decision tree classifiers are excellent in positive predictive value but poor in true positive rate. That is to say, future researchers are encouraged to adopt other models and alter their parameters to test if they perform better in all indicators. Furthermore, they are also encouraged to conduct their research based on more comprehensive and population-specific data, which would theoretically be more accurate in predictions and would be more helpful for organizations in certain regions to detect and predict their employees' mental health state and the general public to test and predict their mental health conditions.

References

1. Galderisi, S., Heinz, A., Kastrup, M., Beezhold, J., & Sartorius, N.: Toward a new definition of mental health. World psychiatry, 14(2), 231 (2015).
2. Harvey, S. B., Henderson, M., Lelliott, P., & Hotopf, M.: Mental health and employment: much work still to be done. The British Journal of Psychiatry, 194(3), 201-203 (2009).
3. Tajalli, P., & Ganbaripanah, A.: The relationship between daily hassles and social support on mental health of university students. Procedia-Social and Behavioral Sciences, 5, 99-103 (2010).

4. Usher, K., Durkin, J., & Bhullar, N.: The COVID-19 pandemic and mental health impacts. International journal of mental health nursing, 29(3), 315 (2020).
5. Moreno, C., Wykes, T., Galderisi, S., Nordentoft, M., Crossley, N., et al.: ow mental health care should change as a consequence of the COVID-19 pandemic. The lancet psychiatry, 7(9), 813-824 (2020).
6. Hamouche, S.: COVID-19 and employees' mental health: stressors, moderators and agenda for organizational actions. Emerald Open Research, 2 (2020).
7. Hofmann, L. A., Lau, S., & Kirchebner, J.: Advantages of machine learning in forensic psychiatric research—uncovering the complexities of aggressive behavior in schizophrenia. Applied Sciences, 12(2), 819 (2022).
8. Sperandei, S.: Understanding logistic regression analysis. Biochemia medica, 24(1), 12-18 (2014).
9. Stoltzfus, J. C.: Logistic regression: a brief primer. Academic emergency medicine, 18(10), 1099-1104 (2011).
10. Westreich, D., Lessler, J., & Funk, M. J.: Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. Journal of clinical epidemiology, 63(8), 826-833 (2010).
11. Friedl, M. A., & Brodley, C. E.: Decision tree classification of land cover from remotely sensed data. Remote sensing of environment, 61(3), 399-409 (1997).
12. Deng, X., Liu, Q., Deng, Y., & Mahadevan, S.: An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Information Sciences, 340, 250-261 (2016).
13. Metz, C. E.: Basic principles of ROC analysis. In Seminars in nuclear medicine, 8(4), 283-298 (1978).