The Research on Depression Prediction Among Student Based on Logistic Regression

Ruge Sun *

Shandong University of Finance and Economics, Jinan, 250002, China

* Corresponding Author Email: lenbe@ldy.edu.rs

Abstract. Depression is one of most common mental illness, which influences human emotions and behavior. High proportion of students now are suffering from depression and predicting depression among students are becoming more and more significant in order to intervention treatment early. The student depression dataset is used to make a predictive model. Firstly, variables are selected by random forests. Next, the top four importance of variables are chosen to fit the logistic regression model. The four variables are having suicidal thoughts, academic pressure, financial stress and age. All of four variables contribute significantly to the model, which p-value are all less than 0.001. After that, the model is used in training data to evaluate performance and is compared with model with all variables. In this study, the AUC of model is good, at 0.905. However, the accuracy, sensitivity and specificity are 0.835, 0.780, 0.874, respectively, which are not high. Therefore, this model is not satisfying to predict depression among students. Multiplicate reasons could lead to low quality of model, such as variables selection. Further research could use other method to do variables selection or use other machine learning models to predict student depression.

Keywords: Depression; random forest; logistic regression.

1. Introduction

Depression is one of the negative emotions, which is manifested as being in low spirits persistently, lack of interest and delight, decline of energy and low self-evaluation. Being depression in the long term even leads to psychological disorders. In extreme cases, people with depression have recurring thoughts of suicide or self-harm [1]. According to the World Health Organization, depression will become a great hazard to humanity in the 21st century and it is the leading factor causing worldwide disparity, affecting approximately 322 million individuals [2, 3]. Recent years, personal and social well-being have been influenced seriously by the prevalence of depression in terms of multiple aspects. Depression has a negative effect on academic performance, decision-making, interpersonal relationship and qualify of an individual's life [4]. The lifetime prevalence of major depression is 16.6% [5]. Currently, depression has aroused global concern. Young people, especially students, are more vulnerable to multiple mental health problems [6].

Depression among college student is a common and serious problem in universities, which is mainly caused by unbalanced development of students' self-consciousness. Yong individuals in university have become more focus on self and have become aware of their limited capacity, resulting in a strong desire to enrich and strengthen themselves. However, it is a complex and difficult process. Interpersonal relationships and pressure caused by academic program are also the reasons of depression among college students [2]. Precious study had demonstrated that a majority of graduate students, approximately 85%, spent more than 41 hours per week on academic program. Additionally, 74% of these students could not finish school on time and 79% of those expressed uncertainty about career prospects and professional future [7]. Evans et al. revealed that the percentage of graduate students suffering depression and anxiety were six times higher than common people [8]. According to Nature in 2019, 36% of PhD students had asked for help resulting from depression and anxiety, which figure was three times higher than that in 2017 [7]. A study conducted by Brownlow et al. in Australia indicated higher degree research (HDR) students, such as those pursuing master degree or PhD, faced a higher risk of mental health issues comparing with others during the COVID-19 lockdown. Approximately 25% of HDR students were at risk of developing mental health problems,

however, the figure of common community was 15.7% [9]. PhD students suffer from unique pressure, which is related with their study experience, academic environment, peer pressure, anxiety about age and gender. Some those mention that they experience periodic minor depression since in university [10]. The study showed that depression is related to quality of sleeping significantly and influence each other. Among patients with depression, approximately 90% of these have sleep disorder, such as insomnia and narcolepsy [11, 12]. Eating sugary drinks, meat and fast food has positive relationship with depression, while the healthy diet has negative relationship with depressive symptoms [13].

Students in universities suffer from the risk of severe depression. It is particularly vital to explore the relationship among depression, GPA, academic stress and satisfaction, educational background and sleep quality. Research demonstrates that depression is related to the changes of internal environmental hormone levels [14]. Teng et al. used random forest method to show that overwork, poor work life balance and teacher-student relationship are significantly correlated with depression [15]. However, there are limited research on age, financial stress, academic pressure or other factors with depression for higher degree students such as class 12 students, undergraduates and postgraduates., using logistic regression. Most studies used random forest model for prediction. This paper aims to use data from these types of students to predict depression with binary logistic regression.

2. Methods

2.1. Data Source

The Student Depression dataset used in this study comes from the Kaggle platform, which is owned by Shodolamu Opeyemi. The original dataset is csv file and it contains 27901 cases and has 3 missing values. The usability calculated by Kaggle is 10.0. These datasets are valuable for research in psychology, data science, and education to identify factors contributing to student depression and to design early intervention strategies.

2.2. Data Introduction and Data Processing

The dataset contains 18 variables. Because the cases are all students, the variables “Work Pressure” and “Job Satisfaction” are all 0 and the variables “Profession” are all Students. The variables “id” is unique identifier for each student and is not related to depression. Therefore, the variables “Work Pressure” “Job Satisfaction” “Profession” and “id” are excluded in analysis of this data. The remaining 14 variables are represented in Table 1.

Table 1. Data Explanation

Variables	Type	Range
Gender	Categorical	0-Male, 1-Female
Age	Numeric	18 to 59
City	Categorical	52 regions
Academic Pressure	Numeric	0 to 5
CGPA	Numeric	0 to 10
Study Satisfaction	Numeric	0 to 5
Sleep Duration	Categorical	0-“5-6 hours”, 1-“7-8 hours”, 2-“Less than 5 hours”, 3-“More than 8 hours”, 4-“Others”
Dietary Habits	Categorical	1-Healthy, 2-Moderate, 3-Unhealthy, 4-Others
Degree	Categorical	28 degrees
Suicidal Thoughts	Categorical	0-No, 1-Yes
Study Hours	Numeric	0 to 12
Financial Stress	Numeric	0 to 5
Family History of Mental Illness	Categorical	0-No, 1-Yes
Depression	Categorical	0-No, 1-Yes

This data has three observations with missing values. The missing values occupy little proportion of whole dataset. Therefore, this paper deletes the observations with missing value directly. There are 27898 observations and 14 variables left in the dataset.

2.3. Method Introduction

The dataset has 13 variables and it is large number for fitting the model. If this paper uses all variables to build the model, the model will be complex. A model that is too complex may overfit the training data, performing well on the training set but poorly on the test set or new data and some variables have lower correlation with depression. Therefore, random forest is used in the paper to do model selection. Mean Decrease Accuracy is used as criterion and the top four variables with the highest this value is chosen to made the logistic regression model. Mean Decrease Accuracy can be represented as:

$$MeanDecreaseAccuracy = \frac{1}{N}\sum_{i=1}^{N}(Accuracy_{original} - Accuracy_{permuted}) \quad (1)$$

Where N is the number of decision trees, $Accuracy_{original}$ is the original accuracy and $Accuracy_{permuted}$ is the accuracy after permuting the variable values.

This paper converts categorical data into factors and uses 80% of the observations as training data. The remaining observations is testing data. Logistic regression is used on selected variables. The logistic regression model can be represented as:

$$P(y=1) = \frac{\exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n)}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n)} \quad (2)$$

Where β_0 is the intercept, $\beta_1, \beta_2, \cdots, \beta_m$ are the coefficients of the independent variables and $x_1, x_2, \cdots, x_n$ are the independent variables. The threshold is 0.5. If $P(y=1) \geq 0.5$, predict the y is 1, which means having depression. If $P(y=1) < 0.5$, predict the y is 0, which means having no depression.

In this paper, the fitted model is used on testing data and confusion matrix and roc curve are used on model evaluation.

3. Results and Discussion

3.1. Descriptive Analysis

Figure 1 is bar plot of suicidal thoughts. Students who have had suicidal thoughts are more tend to have depression and only minority of students who have thoughts of suicide are not depression patients. While for the students who do not have the thoughts of suicide, most of them do not suffer from depression, and only a small number of them suffer from this mental disease.

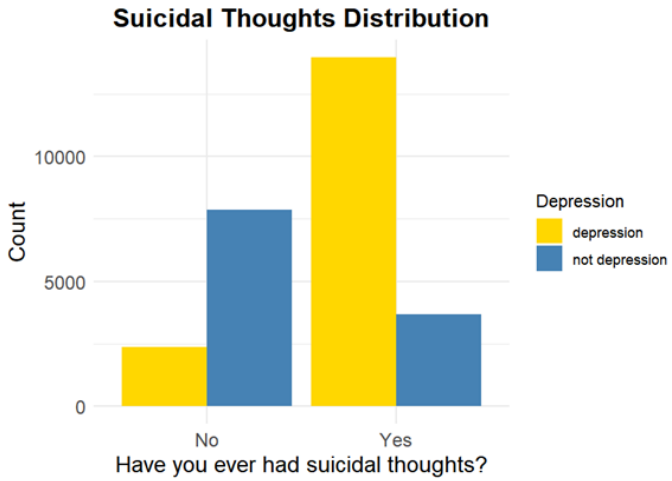

Fig. 1 Suicidal Thoughts Distribution

Figure 2 is boxplot of academic pressure. The mean of academic pressure for depression student is 4, which is much higher than the mean for students having no depression, at 2. The academic pressure scores of depression student are more concentrated at 3-5. However, the scores of students without depression are more likely 1-3. Therefore, students with depression have higher academic pressure than those without depression.

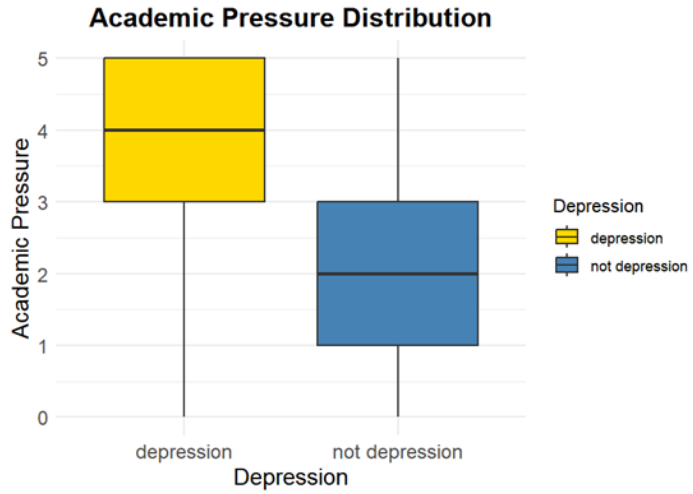

Fig. 2 Academic pressure distribution

Figure 3 is smooth line plot of age. With the increase of age, the probability of suffering from depression also shows a downward trend. Hence, the students have higher age, the probability of depression is lower.

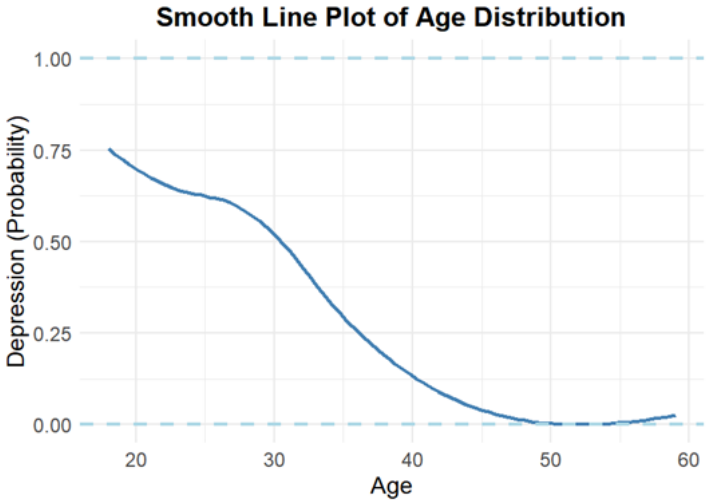

Fig. 3 Smooth Line Plot of Age Distribution

3.2. Variables Selection

Mean Decrease Accuracy is a method to measure variable importance by permuting the values of a variable and observing the change in model accuracy. Figure 4 represents variables importance. This illustrates the importance of various variables in a random forest model based on the Mean Decrease Accuracy metric. The importance of top four variables is “have you ever had suicidal thoughts”, “academic pressure”, “financial stress” and “age”. These variables contribute the most to the model's predictive power, with importance values all exceeding 50.

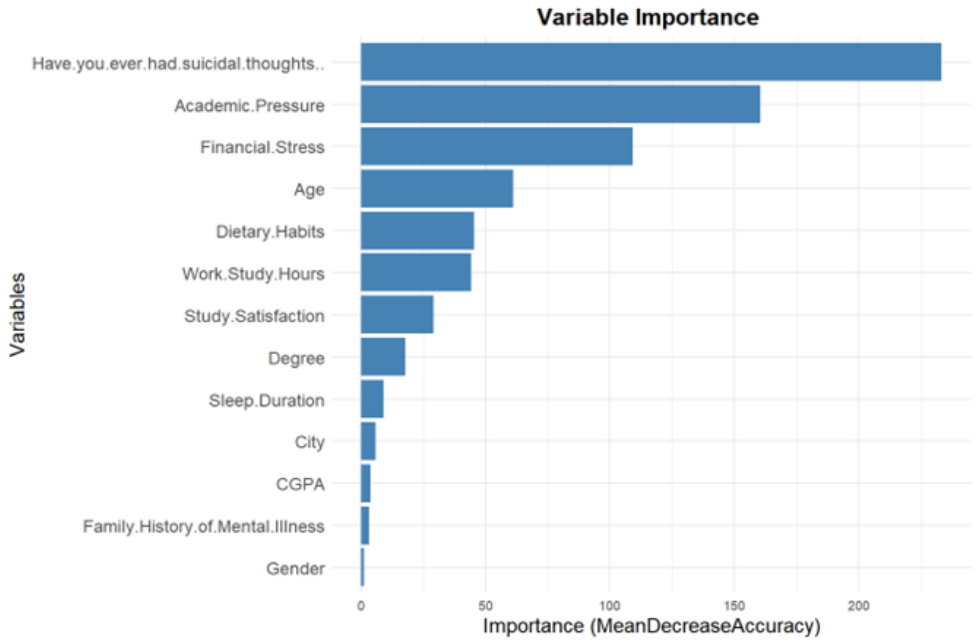

Fig. 4 Variables Importance

3.3. Logistic Regression Results

The variables selected by random forest are used to fit the logistic regression. Table 2 shows the coefficient estimates of logistic regression. The p-values for these four variables are all less than 0.001, indicating that these variables are highly statistically significant in predicting depression. This means they have a notable impact on the occurrence of depression. If one has had suicidal thoughts, the log-odds of having depression will increases by 2.530 compared to those without this thought. Academic pressure, with an estimated value of 0.848, suggests that for each unit increase in academic pressure, the log-odds of having depression will increases by 0.848. Financial stress, with an estimated value of 0.556, indicates that for each unit increase in financial stress, the log-odds of having depression will increases by 0.556. Age, with an estimated value of -0.106, means thar for each unit increase in age, the log-odds of having depression will decreases by 0.106.

Table 2. Logistic Regression Coefficient

Term	Coefficient	P-value
Intercept	-2.728	<0.001
Having suicidal thoughts	2.530	<0.001
Academic Pressure	0.848	<0.001
Financial Stress	0.556	<0.001
Age	-0.106	<0.001

3.4. Model Evaluation

Confusion matrix and Receiver Operating Characteristic (ROC) curve are used to evaluate logistic regression. Figure 5 represents the confusion matrix heatmap, which displays the relationship between the predicted results and actual results of a classification model. The true negative (TN) is 1816, the false positive (FP) is 409, the false negative (FN) is 512, the true positive (TP) is 2843.

Figure 6 represents the ROC Curve. It shows that the Area Under Curve (AUC) value is 0.905 and is close to 1, which shows an excellent ability of the logistic regression model to distinguish between depression and not depression. This also suggests that the model has a relatively high performance in predicting depression.

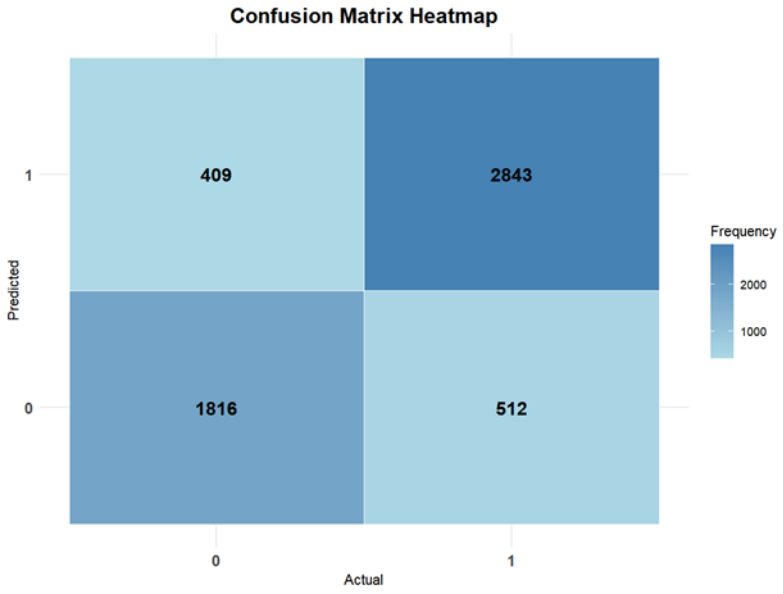

Fig. 5 Confusion Matrix Heatmap

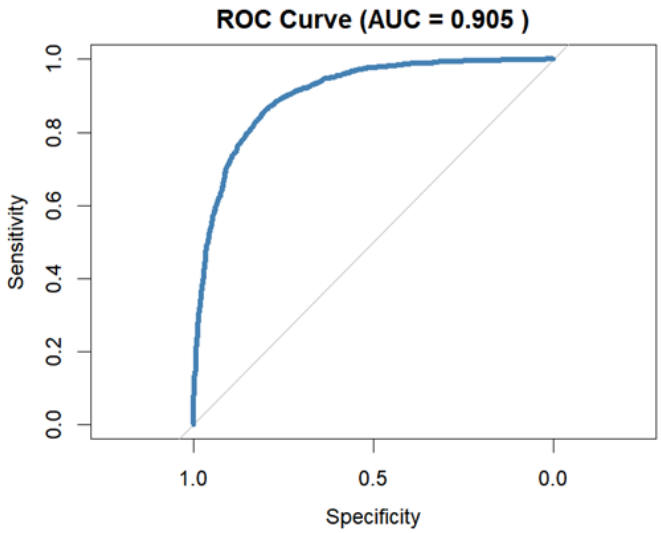

Fig. 6 ROC Curve

Table 3 shows the accuracy, AUC, sensitivity and specificity. The accuracy of the model is 0.835, indicating a good prediction performance. The sensitivity, at 0.780, meaning that the model correctly identifies 78% of the actual positive cases. Meanwhile, the specificity is 0.874, reflecting the model's capacity to correctly identify negative cases.

Table 3. Model Evaluation Metrics

Performance Measures	Value
Accuracy	0.835
AUC	0.905
Sensitivity	0.780
Specificity	0.874

3.5. Comparison with Full Model

Some metrics of model are not perfect, such as sensitivity, which is lower than 0.8. This bad performance may account for variables selection. Therefore, the full model included all of variables was fitted to compare with original regression. And the performance of full model represents in Table 4.

Table 4. Full Model Performance

Performance Measures	Value
Accuracy	0.849
AUC	0.921
Sensitivity	0.799
Specificity	0.885

Although the accuracy, AUC and sensitivity of full model are higher than original one, the difference is slight and the specificity decreases. The performance of full model is not better than the model with variables selection. Therefore, there is no need to use all of variables to predict depression.

4. Conclusion

The variables of having suicidal thoughts, academic pressure, financial pressure and age ware selected by random forests method and are used to build the logistic regression model. Four variables used to fit the model are statistically significant, which p-value are all less than 0.001. For model evaluation, AUC is 0.905, which is really high, indicating that the logistic regression model predicts the testing data well. And accuracy, sensitivity and specificity are good, but not perfect. These three values are all lower than 0.9, especially for sensitivity, which is to evaluate the capability in detecting positive cases. Lower sensitivity indicated higher probability of misdiagnosing students with depression as not having the depression. In the real cases, misdiagnosing a patient as healthy can be extremely dangerous, as it may lead to patient not being identified and treated in time. Therefore, this logistic regression is not so satisfying to predict student depression. However, the performance of model is influenced by various reasons. This paper, only four variables are used in fitting the logistic regression, which may lead to lower accuracy sensitivity and specificity. Therefore, further research could add more variables to the logistic model or using other method to do variables selection such as lasso. Other algorithms could be used to optimize the model as well.

References

[1] Hongyu Zou, Junyao Gao, Wanchun Wu, Lijuan Huo, Wei Zhang. Which comes first? Comorbidity of depression and anxiety symptoms: A cross-lagged network analysis. Social Science & Medicine, 2024.

[2] Shang Youguo. The situation and causes of college students' depression analysis. Journal of North China Institute of Water Resources and Hydropower, 2009, 25(02): 102-104.

[3] Darío Moreno-Agostino, Yu-Tzu Wu, Christina Daskalopoulou, Tasdik Hasan M, Martijn Huisman, Matthew Prina. Global trends in the prevalence and incidence of depression: a systematic review and meta-analysis. Journal of Affective Disorders, 2021.

[4] Zhang J, Peng C, Chen C. Mental health and academic performance of college students: knowledge in the field of mental health, self-control, and learning in college. Acta Psychol, 2024.

[5] Kessler R C, Berglund P, Demler O, Jin R, Merikangas K R, Walters E E. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry, 2005, 62: 593-602.

[6] Garmabi M, Andishmand Z, Naderi F, Sharifnezhad A, Darrudi F, Malekzadeh R, Amini A, Gholami A. The prevalence of depression and anxiety and its association with sleep quality in the first-year medical science students. Depress Res Treat, 2024.

[7] Woolston C. PHD poll reveals fear and joy, contentment and anguish. Nature, 2019.

[8] Evans T M, Bira L, Gastelum J B, Weiss L T, Vanderford N L. Evidence for a mental health crisis in graduate education. Nat Biotechnol, 2018, 36(3): 282-284.

[9] Brownlow C, Eacersall D, Nelson C W, Parsons-Smith R L, Terry P C. Risks to mental health of higher degree by research (HDR) students during a global pandemic. PLoS One, 2022.

[10] Cheng Meng, Li Jiayi. Melancholy at the ivory spire - a narrative study on the depression experience of doctoral students. Educational Research, 2002, 43(07): 88-103.

[11] Jones-White D R, Soria K M, Tower E K B, Horner O G. Factors associated with anxiety and depression among U.S. doctoral students: Evidence from the gradSERU survey. J Am Coll Health, 2022, 70(8): 2433-2444.

[12] Brittany L. Mason, Abram Davidov, Abu Minhajuddin, Madhukar Trivedi H. Focusing on insomnia symptoms to better understand depression: A STAR*D report. Journal of Affective Disorders, 2020.

[13] Mou Xingyue, Tao Shuman, Xie Yang, et al. College students' dietary patterns associated with depressive symptoms. Journal of School Health in China, 2022, 10: 1520-1524.

[14] Ni Min, Wu Qi. Thyroid hormone with depression correlation analysis. Chinese Modern Doctors, 2020, 58(10): 4.

[15] Teng C, Yang C, Liu Q. Utilising AI technique to identify depression risk among doctoral students. Sci Rep, 2024, 14(1): 31978.