UNIVERSITY OF MICHIGAN

STATS415 Data Mining and Statistical Learning

Predicting Depression Condition: Integrating Biochemical and Socioeconomic Indicators

Project Group 3
Ziyang Xiong
Runhui Xu
Xixiao Pan
Huijie Tang

December 3, 2023

Contents

1 Introduction

Depression, a widespread mental disorder, affects approximately 280 million people globally, accounting for nearly 5% of the adult population[1]. This disorder presents significant challenges in diagnosis and treatment, with predictions indicating it could become the leading contributor to the global disease burden by 2030[2]. Therefore, accurate prediction of depression has become a topic of concern for many people. Recent studies have focused on using biochemical markers like gamma-glutamyl transferase, glucose, and triglycerides to gauge depression severity[3]. Parallel to biochemical research, socioeconomic factors have been increasingly recognized for their impact on mental health. Noori Akhtar Danesh and his colleagues highlight the influence of factors such as education level, income, and marital status on depression severity[4], though their full influence remains under-explored.

To this end, this study seeks to answer two critical questions:

1. How can biochemical indicators be used to predict depression severity in mental health diagnoses?
2. To what extent do socioeconomic indicators influence the diagnose of depression?

All in all, the significance of this research lies in its potential to enhance understanding of depression, combining biochemical and socioeconomic predictors. This approach could lead to more accurate, personalized diagnostic tools, improving treatment effectiveness and mental health outcomes. Furthermore, these findings could inform healthcare policies, promoting a more comprehensive approach to mental health care.

2 Data

We explore datasets from the NHANES 2017-March 2020 pre-pandemic period in CDC website, including Depression Score, 9 Biochemical variables, and 5 Social variables.

2.1 Depression Score

To find the severity of depression, we use data from Mental Health - Depression Screener, which records results of the Patient Health Questionnaire (PHQ-9). We denote the sum of PHQ-1 to PHQ-9 as Depression Score and a score ≥ 10 indicates the diagnosis of depression according to reference [5]. Based on the criteria, 13.62% people are diagnosed with depression. The distribution of Depression Score is plotted below.

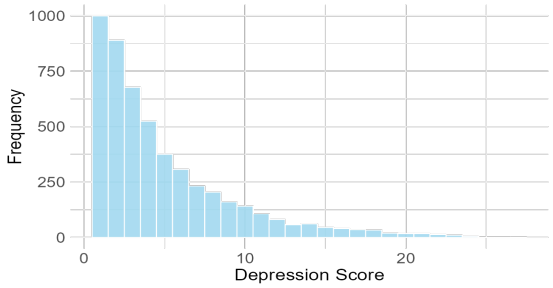

Figure 1: Original Depression Score Distribution

2.2 Biochemical Variables

For the first question, research suggests that there is no simple biochemical indicator diagnosing depression, but indicators in blood and urine may be indicative. Thus, we select several variables from different aspects, which is demonstrated in the table above.

Biochemical Variables	Notation	Source
HDL Cholesterol(mg/dL)	HDL	Cholesterol - High - Density Lipoprotein
C-Reactive Protein(mg/dL)	CRP	High-Sensitivity C-Reactive Protein
Glycohemoglobin(%)	GHB	Glycohemoglobin
Albumin(ug/mL)	ALB	Albumin - Urine
Creatinine(mg/dL)	CRE	Creatinine - Urine
Cotinine(ng/mL)	COT	Cotinine and Hydroxycotinine - Serum
Hydroxycotinine(ng/mL)	HYD	Hydroxycotinine - Serum
Total Cholesterol(mg/dL)	TCH	Cholesterol - Total
White blood cell count	WBC	Complete Blood Count

Table 1: Contents, Notation and data sources of Biochemical Variables.

In Figure 2, we make the scatter plot between Depression Score and biochemical variables, and find that there is no explicit relationship between them.

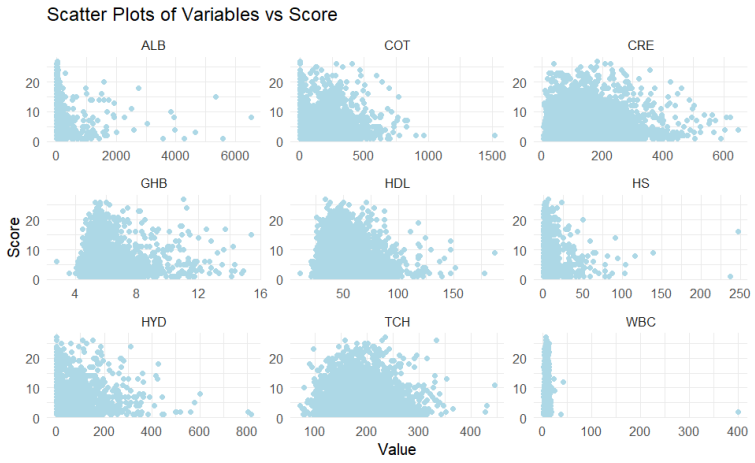

Figure 2: Scatter plots of the relationship of 9 biochemical variables with Score.

2.3 Socialeconomic Variables

For the second question, Table 2 illustrates the socialeconomic variables we use, and detailed meaning of the value of Education variable is shown in Table 3.

Social Variables	Notation	Explanation
A ratio of family income to poverty guideline	RTP	Range from 0-5 and higher value represents less poor
Education level	EDU	5 different levels, the higher the value, the higher the degree. More details are in the next table
Marital status	MAR	1 represents married, 2 represents widowed or divorced and 3 for never married
Gender	Gender	0 represents female and 1 represents male
Age	AGE	value represents years of age

Table 2: Contents, Notation and data sources of Social Variables.

Education Value	Value description
1	Less than 9th grade
2	9-11 grade (Includes 12th grade without diploma)
3	High school graduate/GED or equivalent
4	Some college or AA degree
5	College graduate or above

Table 3: The meaning of the value of Education.

In order to explore whether socioeconomic factors are associated with depression, we use bootstrap method to sample from original data and use frequency of depression population under different situations to estimate the probability of depression. Then we obtain the corresponding 95% confidence interval of depression possibility shown in Figure 3 and Figure 4. We can find obvious tendency in all five parameters, drawing the following conclusions. First, the higher the education level the less possibility to suffer from depression while inverse in the RTP. Second, divorced people are relatively more possible to get depression. Third, people who age between 50 and 60 experience extremely higher depression probability. What's more, female people are more likely to suffer from depression than male ones.

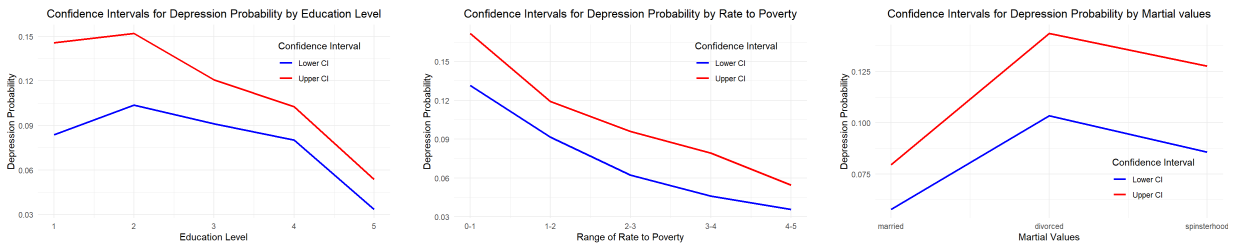

Figure 3: CI for Depression Probability by Education Level, Marital Status, and Rate to Poverty.

**The red line represents the upper value of CI. The blue line represents the lower value of CI.*

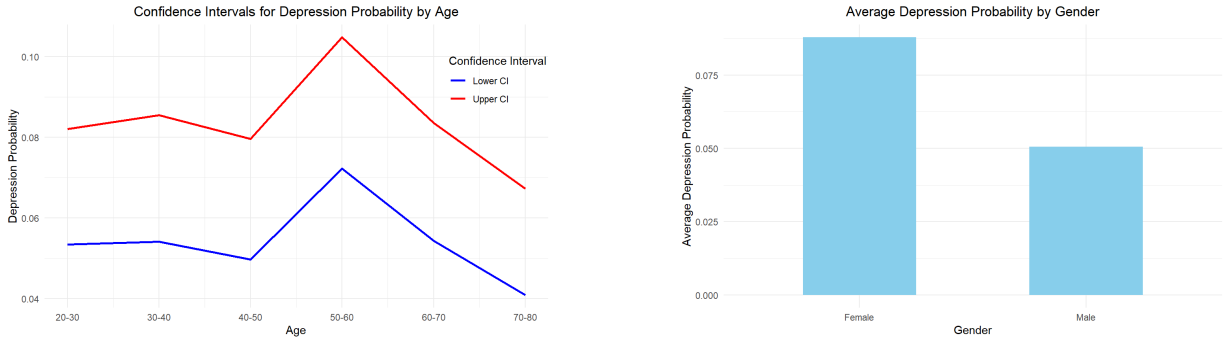

Figure 4: Relationship of Depression Probability with Age and Gender.

**The left figure represents the confidence interval for Depression Probability by Age. The right figure represents the average Depression Probability by Gender.*

3 Methods

3.1 Over-sample method

In analyzing our dataset, it's apparent that depression cases form a minority (depression score $\geq$ 10), posing a challenge for machine learning models which often bias towards predicting the majority class in Figure 1. This skew can lead to deceptively high accuracy rates, masking the model's inefficiency in identifying less common categories. To counteract this, we implemented oversampling in data preprocessing, enhancing the model's capability to accurately detect these less prevalent depression cases. By oversampling observations from minority class (depressed people), we obtained a data set with equal amount of depressed and not depressed observations.

3.2 Modeling Relationship between Depression Score and Biochemical Variables

To predict Depression Score by biochemical variables, first, we apply LASSO regression on oversampled 9 possible biochemical variables to eliminate non-related variables. The model is

$$y = \sum_{j=1}^{p} \beta_j x_j + \beta_0 \tag{1}$$

where y is Depression Score, p is the number of features, x_j is the value of j^{th} variables, and β_j is the coefficients for the j^{th} feature. The estimated coefficients of LASSO regression can be shrunken to zero by minimizing the equation

$$\text{argmin}_{\beta_0,\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2}$$

where λ is the regularization parameter. The regularization term $\lambda \sum_{j=1}^{p} |\beta_j|$ is added to introduce a penalty for large coefficients. We use cross-validation to select the optimal λ, which controls the strength of the penalty. Then, we use the best lambda to generate a LASSO regression and exclude non-related variables.

After eliminating some weak variables by LASSO, we construct a model to predict the Depression Score based on the remaining important factors. The plot between Depression Score and predictors in section 2 suggests that the relationship is non-linear. Therefore, we use Random Forest to make the prediction due to its great flexibility. Leveraging the strength of multiple decision trees, the ensemble structure allows random forest to produce robust and accurate predictions by capturing complex relationships data. Since how biochemical variables reflect level of depression is complicated, random forest is a good choice for our model compared with other prediction model. In addition, we use RMSE to estimate how good our model is, and the formula is illustrated below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

Although random forest improves the accuracy over decision tree at the cost of some interpretability, we can still study different variables' contribution to the model by indicator %IncMSE. It is calculated by randomly permuting values of a specific predictor across all observations, and recording the percentage increase of MSE. Therefore, a larger %IncMSE suggests that the corresponding variable plays a more crucial role in the model.

3.3 Modeling relationship between Depression Score and Socialeconomic Variables

In this section, we intend to make some references about the relationship between depression and socialeconomic variables like marital status, education level, family income, age and gender. Given that marital and gender are categorical data, we use one-hot coding to represent them. Since this is a classification problem, logistical models are first used to give some inferences. Taking family income, marital status, gender, age and education into account, the final logistics classification model is

$$log(\frac{P(X)}{1-P(X)}) = \beta_0 + \beta_1 I + \beta_2 M_2 + \beta_3 M_3 + \beta_4 G_2 + \beta_5 A + \beta_6 E$$

I represents family income, $M_2 = 1$ represents widowed or divorced marital status, $M_3 = 1$ for never married and both 0 for married. $G_2 = 1$ for male and 0 for female. A is people's age and E is education level.

As we assume that depression may happen on people with similar socialeconomic features, it is plausible that these depressed observations will have closer distance with each other in the feature space. Therefore, KNN model seems a good candidate for distinguishing depressed people from others. Our model classifies people into 2 classes, TRUE for depressed people and FALSE for healthy people. The decision algorithm for KNN model is:

$$Pr(Y = j|X = x_0) = \frac{1}{K}\Sigma_{i \in N_0} I(y_i = j)$$

j equals TRUE or FALSE. x_0 represents the vector of the person's social features and Y is the predicted class. N_0 represents the set of k nearest neighbours of x_0 in the feature space and y_i represents their corresponding classes.

4 Results

4.1 Over-sampled Data

We sample 20% data as test set, and use the remaining data as training set. After over-sampling training observations with depression score greater than or equal to 10, the distribution is plotted in Figure 5. Individuals with depression are accentuated, which reduce the bias and enhance overall model performance.

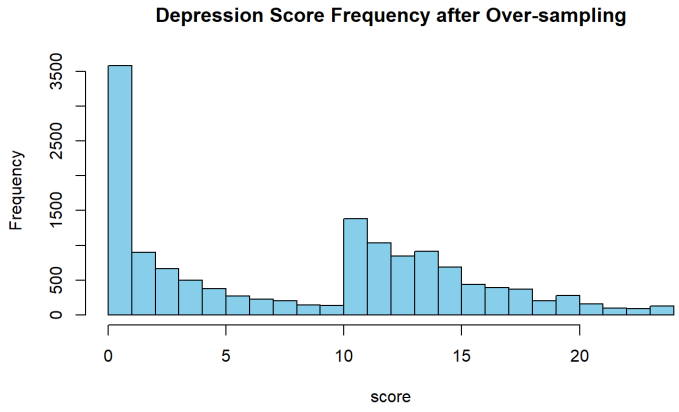

Figure 5: Depression Score Distribution after Over-sampling

4.2 Predicting Depression Score Based on Biochemical Variables

4.2.1 Variable Elimination

After applying cross-validation with 5 folds and the LASSO regression on 9 oversampled biochemical variables, we find that the coefficient of WBC is zero. It means WBC barely has relationship with Depression Score. Thus, we eliminate this variable in the following prediction.

4.2.2 Random Forest Prediction

We construct the random forest model to predict Depression Score based on remaining important variables. The RMSE values are 1.43 and 4.31 respectively on training set and test set. While there is some overfitting, the model demonstrates satisfactory performance on the test data, considering that Depression Score ranges from 0 to 27. Moreover, we compute %IncMSE for each variable, quantifying their respective contributions to the model as illustrated in the following plot.

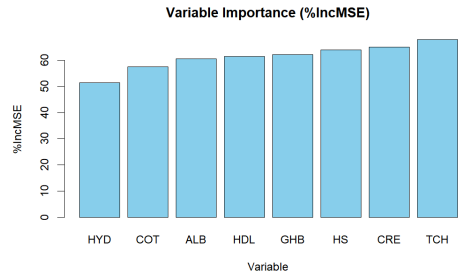

Figure 6: Variable Importance by %IncMSE

The plot reveals that CRP has the most impact, followed by CRE, GHB, and HDL. In comparison, COT and HYD exhibit relatively less importance.

4.3 Potential of predicting depression by using socioeconomic indicators

The final logistic model is $log(\frac{P(X)}{1-P(X)}) = -0.241 - 0.378I + 0.381M_2 + 0.220M_3 + 0.400G_2 - 0.004A - 0.182E$. The significance of predictors in a logistic model is commonly assessed through p-values, and in our case, all the p-values for the included predictors are significantly smaller than 0.02. This observation suggests that all the predictors incorporated into the logistic model are statistically significant and can be considered as influential contributors to the classification.

After applying the KNN model to the oversampled training data, we obtain predicted results and compare them with the observed ones, as shown in Table 4. Based on the data, we are able to identify nearly 70% of depression patients. Unfortunately, there is a 30% chance of diagnosing people incorrectly with the disease. However, it's important to note that we have successfully built a model that can identify depression using socioeconomic factors, which can be valuable information for individuals to learn from.

		Predicted Value	
		False	True
Observed Value	False	812	460
	True	30	68

Table 4: Confusion Matrix of Depression Diagnosis

5 Conclusions and Discussion

When it comes to the first question, we apply over-sample method to reduce the bias of the original data, LASSO regression and cross validation to eliminate useless variables, and random forest to construct the model for diagnosis of Depression Score. We find eight biochemical variables related to cholesterol, protein, nicotine useful in depression prediction. Our random forest model can predict Depression Score with RMSE 4.31 on

test data. It should be noted that RMSE value is much higher on test set than training set, which indicates that there is some overfitting and that our model is not perfect. This may result from the small portion of depression patients in the sample and the fact that although depression would lead to abnormal physical symptoms, it can not be diagnosed solely with biochemical indicators. Nevertheless, our model could still be a reference for mental doctors to consider the severity of depression in patients. Furthermore, in this report, we only explore the relationship of biochemical variables with depression. They may be influenced by other diseases. Besides, in clinical diagnosis, a complete physical examination combined with clinical interviews could better diagnose depression.

When it comes to the second question, the bootstrap analysis from the Data Section distinctly reveals that the confidence intervals for depression probability significantly diverge across varying socioeconomic factors. This may be a convincing evidence of the correlation between socioeconomic status and depression severity. Specifically, a higher level of education and rate to poverty may correlate with a reduced likelihood of depression. This insight guides government to focus mental health resources on lower-income and less-educated demographics. Moreover, our KNN-based model successfully predicts the diagnosis of depression with an accuracy of 64% just by social indicators, offering a potential tool for people to self-diagnose without going to the hospital. However, it's crucial to acknowledge the model is not accurate enough, as it may ignore some patients and falsely identifying others as depressed. The observed inaccuracy in the model could stem from the exclusion of numerous potential socioeconomic factors. Additionally, the limited sample size may also contribute to the model's reduced precision. In the future, people can take other relevant factors into consideration on larger dataset to get better model.

6 Contributions

All members contributed to the topic selection and data processing. Huijie Tang and Xixiao Pan contributed to the first question. Xixiao completed the variable elimination part and Huijie completed the random forest prediction. Runhui Xu and Ziyang Xiong contributed to the second question. Ziyang completed the bootstrap and data visualization and Runhui complete data prepocessing and oversampling. Both Ziyang and Runhui contribute to the model construction. All authors evenly contributed to the corresponding part in code and report writing.

7 Reproducibility

In our report, we identify each data source utilized, all of which are publicly accessible, and provide comprehensive explanations of the specific meanings of each indicator. We thoroughly detail the methodologies implemented for data preprocessing, with corresponding codes provided. Furthermore, the complete suite of codes employed for regularization, visualization, and the construction of predictive models is comprehensively uploaded. The datasets can be found in Github.

References

[1] World Health Organization, “Depression: Fact sheets,” [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression. Accessed on: Nov. 26, 2023

[2] Friedrich M. J., ”Depression Is the Leading Cause of Disability Around the World,” JAMA, vol. 317, no. 15, pp. 1517, 2017. [Online]. Available: https://doi.org/10.1001/jama.2017.3826

[3] X. Li, Y. Mao, S. Zhu, et al., ”Relationship between depressive disorders and biochemical indicators in adult men and women,” BMC Psychiatry, Vol. 23, No. 49, 2023. [Online]. Available: https://doi.org/10.1186/s12888-023-04536-y

[4] N. Akhtar-Danesh and J. Landeen, ”Relation between depression and sociodemographic factors,” Int J Ment Health Syst, Vol. 1, No. 4, 2007. [Online]. Available: https://doi.org/10.1186/1752-4458-1-4

[5] K.Kroenke, R.L. Spitzer, and J.B. Williams, “The PHQ-9: Validity of a Brief Depression Severity Measure,” J. Gen. Intern. Med., vol. 16, no. 9, pp.606-613, 2001, doi: 10.1046/j.1525-1497.2001.016009606.x