Early Detection of Depressed Adolescents with Severe Impairment using Logistic Classifier

[1]Wenhua Lu, Ph.D., [2]I-Ming Chiu, Ph.D., [2]Fangming Tian,
[1]Department of Childhood Studies, [2]Department of Economics, Rutgers University-Camden

Abstract

Using pooled data for adolescents aged 12 to 17 from the annual, cross-sectional National Survey on Drug Use and Health (NSDUH) 2011-2017, this study aimed to build a predictive model to identify severely impaired adolescents with depression using machine learning algorithms.

There are two specific goals of this study:
1) To determine the direction and strength of association between severe functional impairment in adolescents with depression and three potential groups of contributing factors; namely sociodemographic characteristics (i.e., age, ethnicity, income and family structure), parenting style, and school experiences.
2) Based on the findings from the logistic regression in the training data, to construct a predictive model in order to identify depressed adolescents with severe impairment in the test dataset.

Introduction

Major depression, defined as a cluster of specific symptoms with associated impairments, represents a severe mental health concern among adolescents. In the NSDUH, adolescents' 12-month major depressive episodes (MDE) was measured based on the criteria in the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV). MDE-related severe impairment (mdeSI) was measured using the Sheehan Disability Scale, with ratings above 7 on a 0 to 10 visual analog scale.

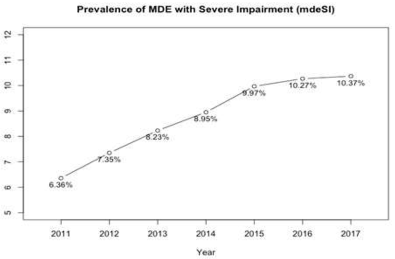


Figure 1 Between year 2011 and 2017, there has been a dramatic increase in the proportion of mdeSI cases.

Methodology and Analysis

Logistic Classifier:

Y is a binary variable with two possible outcomes 1 and 0, and the corresponding probabilities are P and 1 - P, respectively.

$$Y = \begin{cases} 1 \text{ with probability } P, \\ 0 \text{ with probability } 1 - P \end{cases}$$

$$E(Y) = 1{*}P + 0{*}(1 - P) = P$$

The conditional mean function of Y can be written as follows:

$$E(Y|X) = P(Y = 1|X) = \frac{\exp(X * \beta)}{1+\exp(X * \beta)}$$

$$\log\left(\frac{P}{1-P}\right) = X{*}\beta$$

log (Odds) = $\exp(X{*}\beta)$ or $e^{X*\beta}$; exp: exponential function.

Some Terminology about the Confusion Matrix:

Pred \ Actual	0	1
0	TN	FN
1	FP	TP

TP: True Positive TN: True Negative

FN: False Negative FP: False Positive

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$$

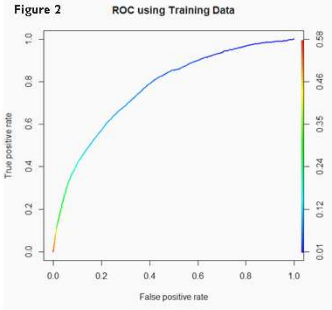
$$\text{Recall} = \frac{TP}{TP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP}$$

Results

Table I Estimation Results (convert estimates to the odds ratio)

(Intercept)	**0.0148****	**0.0131**	**0.0166**
Gender (*ref: Male*)	4.1413**	3.8926	4.4082
Age: 14-15 (*ref: 12-13*)	1.8741**	1.7323	2.0286
Age: 16-17	2.1607**	1.9988	2.3373
Race: Hispanic (*ref: White*)	1.0051	0.932	1.0833
Race: Black	0.6969**	0.6308	0.7689
Race: Asian/NHPIs	0.8433*	0.724	0.9776
Race: Others	1.1522*	1.0331	1.2827
Insurance (*ref: No*)	0.8641*	0.7526	0.9884
Income: $20,000 - 49,999 (*ref: <$20,000*)	1.0576	0.9706	1.1529
Income: $50,000 - 74,999	1.0887	0.9847	1.2037
Income: $75,000 or more	0.9741	0.8875	1.0696
Father at home (*ref: Yes*)	1.1035**	1.0311	1.1806
Mother at home (*ref: Yes*)	1.0807	0.9797	1.1903
Sibling <18 (*ref: Yes*)	1.0464	0.9857	1.1106
Authoritative parenting (*ref: low level*)	2.3519**	2.2042	2.5088
School experiences (*ref: good experiences*)	3.0238**	2.8453	3.213

Figure 2



Final Decision for Threshold Value:

$\hat{P} \geq 0.05$ (by Maximizing Recall)
Training Data (size = 73,764)

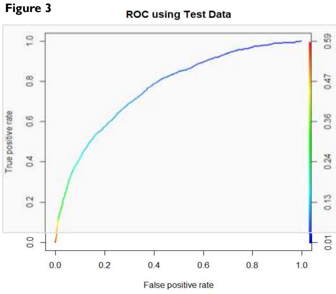
Pred \ Actual	0	1
0	33461	997
1	33976	5330

$$\text{Accuracy} = \frac{33461+5330}{73764} = 52.59\%$$

$$\text{Recall} = \frac{5330}{997+5330} = 84.24\%$$

$$\text{Precision} = \frac{5330}{33976+5330} = 13.56\%$$

Figure 3



Test Data (size = 24,588)

Pred \ Actual	0	1
0	10951	331
1	11521	1785

$$\text{Accuracy} = \frac{10951+178}{24588} = 51.80\%$$

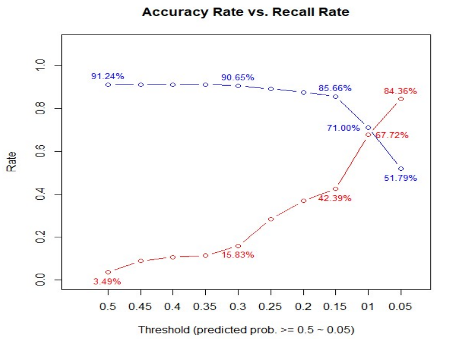
$$\text{Recall} = \frac{1785}{331+1785} = 84.36\%$$

$$\text{Precision} = \frac{1785}{11521+1785} = 13.42\%$$

Figures 2&3. The right scale in both Receiver Operating Characteristic (ROC) cure figures is the range of threshold values. Both true positive and false positive rates increase when the threshold values dropped from the red region (0.47~0.59) to the blue region (0.01~0.13).

Results

Figure 4. When the threshold value is decreasing from 0.5 to 0.05, the chance to detect adolescents with mdeSI increases dramatically (red line) at the cost of lower accuracy rate (blue line)



Conclusion

The performance of the predictive model depends on the decision of the threshold values (i.e. a range of probabilities to identify mdeSI cases in the test data set). When the threshold of the predicted probability is reduced to 0.05, we are able to identify 1,785 out of 2,116 (84.36% of the Recall rate) depressed adolescents with severe impairment. However, this better predictive performance comes with a cost of larger false positive rate (11,521 out of 13,306 cases or 86.58%).

While our empirical results reveal the statistical relationships between the covariates and depression-related severe impairment, the capability of our predictive model is subject to the selection of the threshold value. Further, without specifying an explicit cost or loss function, it would be difficult to choose an optimal threshold value. Nevertheless, once a better predictive model is decided, intervention plans can be developed for those high-risk adolescent groups. Findings from this study, and our continued work, therefore, hold implications to eventually reduce both social and economic cost of depression in adolescents.