Xinyan Deng,
E-mail: edeng@madeira.org

DETECTING NEED FOR SPECIAL EDUCATION: DEVELOPMENT AND VALIDATION OF A PREDICTIVE MODEL

Abstract. Special education is a way of educating students while addressing their individual needs. Circumstances qualified for special education include both physical disabilities and mental illnesses that would impair their learning abilities. Special education makes learning more accessible for students with disabilities. In this report, 18,054 observations from the dataset of 2017 National Health Interview Survey were analyzed. Factors such as gender, race, age, and parents' education levels were examined to develop and validate a model for predicting the probability of a student's need for special education. The predictive model was further validated through the ROC curves. The model demonstrated diagnostic ability through the curves and was created in hopes of detecting need for early intervention services for students so that they can get the proper help they require promptly.

Keywords: special education, predictive model, ROC curves, logistic regression, sort students.

1. Introduction

On February 14, 2018, 17 students and faculty members in Stoneman Douglas High School, Florida died from a massive shooting [6]. 17 others were injured. This deadliest high school shooting in America was committed by 19-year-old Nikolas Cruz, a previous student of the school who had suffered from mental health issues. Despite his mental illness, he attended a regular public school throughout his teenage years, causing many troubles with his peers at school, eventually leading up to the violent shooting. This situation may have turned out differently if Cruz had received special education.

Special education is a way of educating students while addressing their individual needs. Circumstances qualified for special education include both physical disabilities and mental illnesses that would impair their learning abilities [1]. There are many institutions in the United States that provide special education for those in need, and they help students in ways that acknowledge their disabilities. According to a study, there is a rising trend of students receiving special ed, with approximately 13 percent of all students getting such instructions [5].

The hypothesis of this study is that “the likelihood that a person receives special education is related to one or more factors such as his or her race, gender, age, parents' education levels etc.” The objective of this study is to develop a predictive model to detect the relationship between special education and other factors. With this model, schools will be able to collect survey data and approximate the students' likelihood of needing special education.

2. Method

2.1 Data

The dataset of 2017 National Health Interview Survey (herein after referred to as NHIS dataset) was used to identify how different factors have resulted in the need for special education. 18,054 observations were used in this study. Observations with missing data points were excluded from the analysis.

The National Health Interview Survey is a survey done every year since 1957 through personal household interviews by the U.S. Census Bureau [3]. It is one of the major data collection programs of the National Center for Health Statistics (NCHS), which is a part of the Centers for Disease Control and Prevention (CDC). The results help track national health

status. The questions in the survey include investigation of the surveyees' basic background information and health status, such as number of people in the family, occupation of parents, ethnicity, physical disabilities, mental health issues etc.

The NHIS dataset was divided into two datasets: the training dataset (50%) for developing the model and the test dataset (50%) for validating the model.

2.2 Statistical Analysis

The statistical analysis included three stages. In the first stage, the variables that were relevant to the hypothesis were selected from the dataset. In the second stage, a logistic regression model and a neural network model were constructed to analyze the relevance of the factors to each other. This process was conducted through R. In the third stage, corresponding ROC curves were constructed for validation of the predictive model.

2.2.1 Variables:

The outcome variable is whether the person is receiving special education or early intervention services (PSPEDEIS: Does – receive Special Education or EIS?)

Table 1. – Variables used in the study

SEX	1: male 2: female
1	2
RACRECI3	1: White 2: Black 3: Asian 4: All other race groups*
AGE_P	Age 00: Under 1 year 01–84: 1–84 years 85: 85+ years
ORIGIN_I	Hispanic Ethnicity: 1: Yes 2: No
MOM_ED	01 Less than/equal to 8th grade 02 9–12th grade, no school diploma

1	2
MOM_ED	03 School graduate/GED recipient 04 Some college, no degree 05 AA degree, technical or vocational 06 AA degree, academic program 07 Bachelor's degree 08 Master's, professional, or doctoral degree
DAD_ED	01 Less than/equal to 8th grade 02 9–12th grade, no school diploma 03 School graduate/GED recipient 04 Some college, no degree 05 AA degree, technical or vocational 06 AA degree, academic program 07 Bachelor's degree 08 Master's, professional, or doctoral degree

2.2.2 Logistic Regression in R:

Logistic regression models were used in this study to calculate the predicted possibility. Logistic regression is a part of a category of statistical models called generalized linear models. It allows one to predict a discrete outcome from a set of variables that may be continuous, discrete, dichotomous, or a combination of these. Typically, the dependent variable is dichotomous and the independent variables are either categorical or continuous. In this study, the dependent variable is special education; the independent variables are the person's gender, race, age, and parents' education.

The logistic regression model can be expressed with the formula:

$$\ln(P/P-1) = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \dots + \beta_n \times X_n$$

where P is the probability of needing special education, β_0 is a constant, β_1 through β_n are the regression coefficients and X_1 through X_n are the independent variables, such as age, sex, race, parents' education etc. For simplicity, the left-hand side of the equation

is often referred to as “the logit.” The interpretation of the coefficients describes the independent variable’s effect on the natural logarithm of the odds, rather than directly on the probability P . To facilitate interpretation, $e\beta_n$, a transformation of the original regression coefficient β_n can be derived, which can be interpreted as follows:

If $e\beta_n > 1$, $P/(1-P)$ increases.

If $e\beta_n < 1$, $P/(1-P)$ decreases.

If $e\beta_n = 1$, $P/(1-P)$ stays unaffected.

In this study, the cutoff value of P was 0.01. Independent factors with P values lower than 0.01 were considered related to the dependent variable; factors with P value higher than 0.01 were considered not quite related to the dependent variable.

2.2.3 Neural Network in R:

Artificial neural network is made up of information units that are related to each other [2]. It imitates the human nervous system. Different from logistic regressions, neural networks are non-linear statistical data modeling tools. They create different layers to analyze the complex relationship between the input and output variables. Packages called “neuralnet” in R were used to conduct neural network analysis. The package “neuralnet” focuses on multi-layer perceptrons, which are well applicable when modeling functional relationships.

2.2.4 Validation of the Model

After the construction of the models, ROC curves were created to determine the accuracy of the model. Receiver Operating Characteristic curves exhibit the diagnostic ability of a binary classifier system through their area under curve (AUC). The numbers exhibit the percentage of the data that follows the model, ranging from 0.5 to 1, 1 being the score of a perfect predictor and 0.5 being the score of a predictor that makes random guesses. To draw a ROC curve, the

true positive rate (TPR) and false positive rate (FPR) are needed (as functions of some classifier parameter). The TPR defines how many correct positive results occur among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test.

3. Results

3.1 Model Development

3.1.1 Logistic Regression of the Training Dataset

The significance level of the analysis was 0.01. The dependent variable of the model was measured by the question in NHIS survey “Does – receive Special Education or EIS?”, coded as 1 = yes, 0 = no. The factors constructing age, sex (coded as 1 = female, 0 = male), whether the student was Hispanic (coded as 1 = Hispanic, 0 = no), race (coded as 1 = white, 2 = black, 3 = Asian, 4 = other race groups), mother’s education (coded as 1 = less than/equal to 8th grade, 2 = 9–12th grade, no school diploma, 3 = school graduate/GED recipient, 4 = some college, no degree, 5 = AA degree, technical or vocational, 6 = AA degree, academic program, 7 = bachelor’s degree, 8 = master’s, professional, or doctoral degree), father’s education (same coding system as mother’s education) were used as predictors in the logistic regression model. There were 78.132 observations in the original dataset. However, 60.078 observations with missing values were excluded, leaving 18.054 observations in the resulting analysis dataset. The analysis dataset was then split into two datasets in the ratio of 50:50. 50% of the data was used for model development (9.027 observations), and 50% of the data was used for model validation (9.027 observations). The results of the logistic regression analysis of people who receive early intervention services of the training dataset are listed in (Table 2).

Table 2. – Logistic regression for having special education or early intervention services

	Estimate	Std. Error	z value	Pr(> z)	
1	2	3	4	5	6
(Intercept)	–2.338	0.223	–10.467	0.000	***

1	2	3	4	5	6
SEX	-0.678	0.056	-12.056	0.000	***
ORIGIN_I	0.307	0.068	4.491	0.000	***
AGE_P	0.048	0.005	9.076	0.000	***
MOM_college	-0.248	0.068	-3.636	0.000	***
DAD_college	-0.182	0.074	-2.466	0.014	*
White	0.119	0.172	0.688	0.492	
Black	0.083	0.182	0.454	0.650	
Asian	-0.787	0.230	-3.417	0.001	***

About 8.6% of 18,054 school students had special education or early intervention services. Because the cutoff value for P is 0.01, any factors with P value lower than 0.01 were considered relevant.

It can be inferred from the logistic regression table that a person's need for special education is related to the person's gender, Hispanic ethnicity, age, parents' educational status, and race.

Figure 1. Matrix of correlations between variables

Basically, a corrgram is a graphical representation of the cells of a matrix of correlations. The idea is to display the pattern of correlations in terms of their signs and magnitudes using visual thinning and correlation-based variable ordering. Moreover, the cells of the matrix can be shaded or colored to

show the correlation value. The positive correlations are shown in blue, while the negative correlations are shown in red; the darker the hue, the greater the magnitude of the correlation.

According to the corrgram, females were less likely to have special education or early interven-

tion services than males were. Non-Hispanic children were more likely to have special education or early intervention services than Hispanic children were. Older children were more likely too than the younger children were. Asian children and children with more educated parents were less likely to have special education or early interventions services.

3.1.2 Neural Network of the Training Dataset

In (Figure 3), line thickness represents weight magnitude and line color represents weight sign (black = positive, grey = negative). The net is essentially a black box so it is insufficient to indicate fitting, the weights, and the model. However, it can be inferred that the training algorithm has converged, and therefore the model is ready to be used.

Figure 2. Artificial neural network in training sample

Figure 3. Variable importance in artificial neural network

according to this neural network, the most important predictors were age, followed by Asian,

Black, White, and Hispanic origin, sex, and parent's education levels.

3.2 Model Validation – the ROC Curves

According to Figure 4, for training sample, the AUC-ROC score is 0.64 for logistic regression and 0.69 for artificial neural network, meaning that 64

percent of the data in the training dataset follows the logistic regression model and 69 percent follows the neural network model.

Figure 4. Roc in training sample for logistic regression (red) vs neural network (blue)

Figure 5. Roc in testing sample for logistic regression (red) vs neural network (blue)

According to Figure 5, in testing sample, the AUC-ROC score is 0.63 for logistic regression and 0.65 for artificial neural network, meaning that 63

percent of the data in the test dataset follows the logistic regression model and 65 percent of the data follows the neural network model.

4. Discussion

4.1 Model Analysis and Interpretation of the Results

According to the results, out of the 9,027 data points in the training dataset, 782 received special education service, and 8,245 did not. According to the logistic analysis results listed in the table, the predictive model of need for special education is:

Predicted logit of special education needs = $-2.338 - 0.678 \times \text{Gender} + 0.307 \times \text{Hispanic Ethnicity} + 0.048 \times \text{Age} - 0.248 \times \text{Mother's education} - 0.182 \times \text{Father's education} + 0.119 \times \text{White} + 0.083 \times \text{Black} - 0.787 \times \text{Asian}$.

The coefficients of the parameters were interpreted as follows. At the significance level of 0.01:

- On average, controlling other variables, a female student is 49.2 percent less likely to need special education than a male student is.
- On average, controlling other variables, a non-Hispanic student is 35.9 percent more likely to need special education than a Hispanic student is.
- On average, controlling other variables, an older student is 4.9 percent more likely to need special education than a younger student is.
- On average, controlling other variables, a student with mother who received college education is 22.0 percent less likely to need special education.
- On average, controlling other variables, a student with father who received college education is 16.6 percent less likely to need special education.
- On average, controlling other variables, a person who is white is 12.6 percent more likely to need special education.
- On average, controlling other variables, a person who is black is 8.7 percent more likely to need special education.
- On average, controlling other variables, a person who is Asian is 54.5 percent less likely to need special education.

It is indicated in the model that being male, non-Hispanic, older, non-Asian, and having parents with lower education levels are significant predictors in

predicting the need for special education. Schools can pay more attention to these groups of people to ensure that their needs are being met.

According to the artificial neural network (Figure 2) and the results of the ROC curves, it is undeniable that the model is not perfect and errors exist. However, given that the variables of the model were all the most basic background information of a student, the errors can be comprehended and can be decreased by specific analysis of each case.

4.2 Comparison with Previous Studies and Possible Future Development

This study developed a prediction model that relies on a student's most basic background information. In the past, most studies that attempted to predict the need for special education were based on experimental research [4]. Those studies required the presence of the students, making it difficult to predict for a larger group. However, with the prediction model derived from this study, schools would be able to predict the possibility that a student would need special education in a larger scale.

For future study, factors such as students' behavioral characteristics, social relationship with family and schoolmates, history of being abused, and stressful events such as loss of a parent or friend may be included in model development and validation. These factors would improve the precision of the model, because they can narrow down the groups of students who need special education.

5. Conclusion

In this study, logistic regression and neural network were used to develop a predictive model for special education needs among students. Factors like gender, race, age, and parents' education were considered influential factors on this issue. The dataset was split into two to validate the model. The predictive model was validated by ROC curves. The AUC-ROC score of the model was approximately 0.64, meaning that about 64% of the dataset follows the model. This model can be used for schools to evaluate the students' probability of needing special

education to address their learning disabilities. Application of this model can help schools sort students into special education institutions according to their needs for their best future.

References:

1. Association for Children's Mental Health. (n.d.). Education. Retrieved February 24, 2019. From The Association for Children's Mental Health website: URL: <http://www.acmh-mi.org/get-information/child-and-family-services/education>
2. Guest Blog. (2017, September 7). Creating & Visualizing Neural Network in R. Retrieved March 8, 2019. From Analytics Vidhya website: URL: <https://www.analyticsvidhya.com/blog/2017/09/creating-visualizing-neural-network-in-r>
3. National Center for Health Statistics. (n.d.). National Health Interview Survey. Retrieved January 21, 2019. From NHIS website: URL: <https://www.cdc.gov/nchs/nhis/index.htm>
4. Putman J. W., Spiegel A. N., & Bruininks R. H. (1995, May). Future Directions in Education and Inclusion of Students with Disabilities: A Delphi Investigation. Retrieved February 25, 2019. From University of Nebraska website: URL: <https://pdfs.semanticscholar.org/4898/e7020c59ba7228e14349922726fdd1fb15ec.pdf>
5. Salem T. (2018, June 6). Special Education Students On the Rise. Retrieved January 24, 2019. From U. S. Newswebsite: URL: <https://www.usnews.com/news/education-news/articles/2018-06-06/special-education-students-on-the-rise>
6. Washington Post staff. (2018, March 10). Red flags: The troubled path of accused Parkland shooter Nikolas Cruz. Retrieved January 24, 2019. From The Washington Post website: URL: https://www.washingtonpost.com/graphics/2018/national/timeline-parkland-shooter-nikolas-cruz/?noredirect=on&utm_term=.7231d289c056