Machine Learning Models for Detecting Depressive Disorders

Minhwa Lee
The College of Wooster

Leave Empty

This space will be automatically filled with a QR code and number for easy sharing

INTRODUCTION

Depressive disorders (or depression) are the most common but serious health concerns for all age groups in the U.S. In this research, we aim to examine and predict which societal attributes, such as socio-demography and healthcare accessibility, impact the onset of depressive disorders in different age groups of the U.S. residents in 2018, by using three machine learning methods – **decision trees, logistic regression, and support vector machines.** Finally, we compare the performance of each of the three methods in terms of accuracy, precision, recall, and AUC values.

DATA DESCRIPTION

Dataset: 2018 Behavioral Risk Factors Surveillance System (BRFSS)
(Published by the Center of Disease Control and Prevention (CDC))

- Young Group (18 – 40) : 58,113 entries
- Middle Group (41 – 60) : 85,470 entries
- Old Group (61 – 85): 94,625 entries

Selected Variables

Type	Category	Variables
Dependent	Health Records	Depressive Disorders (Ever Diagnosed? Yes or No)
Independent	Socio-Demography	Race Current Employment Status Weight # of Children Marital Status Veteran Status Healthcare Inaccessibility due to High-Medical Cost
	Health States	The number of bad mental days Decision making issues Bone/Muscle-related Illnesses Smoking/Drinking Status Lung Illnesses Sleeping hours
	Health Records	HPV/HIV Test The place that received Flu Shot Intestine Exams Records

Table 1. Selected Variables

METHODOLOGY

1. Decision Trees with CART algorithm

- Enables Binary Decision (Yes or No) for classification tasks
- A white-box model
- Gini Impurity (GI)

For i entries of each J classes (categories) at the node N, we compute the GI of each label N as follows:

$$I_G(N) = \sum_{i=1}^{J} p_i p_j = \sum_{i=1}^{J} p_i(1-p_i) = \sum_{i=1}^{J}(p_i - p_i^2) = 1 - \sum_{i=1}^{J} p_i^2 \quad \text{(Eq. 1)}$$

The lowest GI of the corresponding class J will be placed at N.

2. Logistic Regression

- A white-box model
- Enables quantitative approaches to the effect of each independent variable on the causality of depressive disorders
- A suitable regression method for binary dependent variable
- Check p-values to examine statistically significant variables

For k independent variables $x_1, x_2, \ldots, x_k$ that predict a binary dependent variable Y, the logistic regression model is as follows:

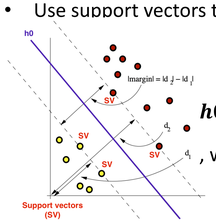
$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2 + \ \ldots + \widehat{\beta_k}x_k, \quad \text{(Eq. 2)}$$

where $\hat{\pi}$ is the predicted probability (odds) of Y = Yes, and $\widehat{\beta_i}$ is the estimated coefficient of each independent variable by the model.

$$\therefore \hat{\pi} = \frac{e^{\{\widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2 + \ \ldots + \widehat{\beta_k}x_k\}}}{1 + e^{\{\widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2 + \ \ldots + \widehat{\beta_k}x_k\}}} \quad \text{(Eq. 3)}$$

3. Support Vector Machine with Gaussian Radial Basis Function (RBF)

- A black-box model
- Separate the region of the data into different groups of points, with using Gaussian RBF function (called "kernel")
- Use support vectors to conduct classifications and predictions



$$h0\text{: } \vec{w} * \vec{x} + b = 0 \quad \text{(Eq. 4)}$$

Maximize margin of h0 $= ||d_2 - d_1|| = \frac{2}{||w||}$

$$\boldsymbol{h0}\text{: } \boldsymbol{f(\vec{x_j}, \vec{\alpha^*}, b^*) = (\sum_{i=1}^{m} y_i [\vec{\alpha^*}]_i K(\vec{x_i}, \vec{x_j})) + b^*} \quad \text{(Eq. 5)}$$

, where $\boldsymbol{K(\vec{x_i}, \vec{x_j}) = \exp(-\gamma \left|\left|x_i - x_j\right|\right|^2)}$ (RBF kernel)

RESULTS

Age Group	Most Influential Attributes
Young	HPV Test Records (Yes) Current Employment Status (Non-wage)
Middle	Lesser monthly alcohols HIV Test Records
Old	Sleeping hours (↓) Marital Status (Not married) Healthcare inaccessibility Due to Medical Cost Issues (Yes)

Table 2. Decision Trees

Age Group	Most Influential Attributes
Young	HPV Test Records (Yes – 118% ↑) Smoking (Yes – 63% ↑)
Middle	Employment Status (Unemployed – 114% ↑) Marital Status (Married – 34% ↓) Flu Shot Place (Non-medical places – 15% ↑)
Old	Healthcare Inaccessibility due to High Medical Cost (Yes – 64% ↑) Intestine Exams (Yes – 51% ↑) Lung-related illnesses (Yes – 71% ↑) Veteran Status (Yes – 26% ↓)

Table 3. Logistic Regression

Methods	Accuracy	Precision	Recall	AUC
Decision Trees	**85.0%**	62%	35%	**65.3 %**
Logistic Regression	84.6%	76%	**65.3%**	62.0%
Support Vector Machine (RBF kernel)	84.5%	**78.3%**	59.3%	59.2%

Table 4. Average Model Performance

DISCUSSIONS

Table 2

- Mental-related attributes (e.g. **the number of bad mental days, decision-making issues**) and **physical disability** are found to be influential in all age groups of the US adult residents.
- The factors related with sexually transmitted diseases (e.g. **HIV/HPV test records**) are influential for young and middle groups.
- **Lesser sleeping hours** and **healthcare inaccessibility due to economic hardship** are influential for the old group.

Table 3

- Having HPV test records leads to **118% higher probability** of the onset of the depressive disorders for young group.
- Healthcare inaccessibility due to high medical costs leads to **64% higher probability** of the onset of the depressive disorders for the old group.

Table 4

- Decision trees excel the others in terms of accuracy, but **SVM with RBF kernel excels in terms of precision.**
- Logistic regression excels the others in terms of recall.

ACKNOWLEDGEMENT

I would like to thank my advisors, Dr. Marian Frazier and Dr. Sofia Visa, for their guidance and support.

Will be covered by controls if you define slides