Identify Risk Factors for Early-onset Depression: A Machine Learning Model

Jiaxu Han

Models

In this project, the goal is to identify the risk factors for Early-onset depression and build a decision tree using the identified risk factors. The decision tree built here is not meant to be exhaustive but to be simple in a sense that clinicians are able to use the tree to quickly identify children at a higher risk for depression with just a few questions. Due to the unbalanced nature of the data set, I decided to train a decision tree and two additional ensemble methods — bagging, and random forest — that are generally considered to be better performed than other types of classifiers on datasets with the unbalanced outcome variable. A simple decision tree determines the predictive value based on a series of conditions. In "Bagging", each decision tree is built based on each of the bootstrapped subsamples and then the algorithm aggregates over all the decision trees to form the final prediction. Random forest is similar to the "Bagging" model, but it splits on a random selection of features each time.

Data

The data used in this project is collected through the National Survey of Children's Health (NSCH). The NSCH collects information about the physical and emotional health of children between 0 to 17 years old. Questions in the survey include access to - and quality of - health care, family structures, and environment, parental health, neighborhood characteristics, as well as school and after-school experiences (US Census Bureau, 2019). The data was collected in 2003, 2007, 2011/12, 2016 and 2017. A previous version of the NSCH survey and a telephone methodology were used in the first three waves of data collections. A revised version of the

NSCH survey was used for data collection in 2016 and 2017 and the survey was distributed and collected via mail and online platform. The NSCH data are available to the public. The current project would only use 2016 and 2017 data since they are the newest and used the revised survey. The 2016 dataset can be found and downloaded on the website: https://www.census.gov/data/datasets/2016/demo/nsch/nsch2016.html. The 2017 dataset can be found and downloaded on the website: https://www.census.gov/data/datasets/2017/demo/nsch/nsch2017.html.

Data Preprocessing and Summary Statistics

Preprocessing

After merging 2016 NSCH topical datasets and implicate dataset, the sample size of the dataset is 50212 with 429 columns. The sample size of the 2017 dataset is 21599 with 431 columns. Though NSCH used revised survey in both 2016 and 2017, there are slight differences in those surveys. Each survey was assigned with a unique household ID (HHID), so there is hardly a way to know if there are households that filled out surveys in both years. Since the surveys used in 2016 and 2017 datasets are almost identical and the time of collection is very close to each other, I decided to merge the two years dataset into one dataset by combining common columns and overlooked the factor of the survey year. After combining both years dataset, the sample size becomes 71811 with 421 columns.

Another issue associated with this dataset is that there are three different types of forms (T1, T2, and T3) designed for 0-5, 6-11, and 12-17 age groups respectively. Those three forms share many common questions but they also got their own questions. It would be ideal to do data analysis for each age group, but the number of respondents reported to have depressed children

largely decreased in each group. To incorporate as many individuals as possible into data analysis, I decided to only consider common variables of those three forms.

In addition, some questions in T1 associated with developmental milestones are specific for children under 3 years old. Since depression among preschool children is rarely reported in the literature review, I screened out children who are lower than 3 years old in the current dataset as well.

Outcome variable

There are three questions associated with the child's depression in the survey: 'K2Q32A' asks whether a doctor or other health care provider has ever told the respondent that his/her child has depression; 'K2Q32B' asks the respondent whether his/her child currently has depression; and 'K2Q32C' asks the severity of the depression. Here I chose whether or not a child has been diagnosed with depression as the main outcome variable. Using one item to decide if a child has depression or not is not ideal. Different children may have different doctors or psychologists and they use different diagnostic criteria. However, a parent's report on a child's mental health is the closest item in this survey to the theoretical constructs.

Selecting Relevant Variables

The dataset has an unbalanced outcome variable with 1 out of 219 reported child having depression. In addition, some columns have a very low response rate. Therefore, it takes a more stringent approach to drop rows and columns. I first manually selected 112 variables to drop from the main dataset. Those variables have not been considered relevant to early-onset depression based on a literature review including information about breastfeeding and why does one have or not have insurance and etc. After re-coding some responses, I dropped columns with

the response rate lower than 80% and then dropped rows with missing values. As a result, a sample size of 32896 and 237 columns are included for data analysis. Now, the majority of cases versus minority cases ratio becomes around 29:1.

Descriptive Statistics

In the cleaned dataset, the mean age of the children is 10.45, ranging from 3 to 17. 80.9% of the children are white, 3.9% are black, and 5.0% are Asian. Only 4.3% of the respondents reported a child having depression. 76.9% of the children who had been diagnosed with depression still have depression at the time the survey was filled out.

Methods and Initial Results

Hyperparameter Tuning and Model Evaluation

In order to find the suitable parameter grid for each of the three models, I used the built-in method “GridSearchCV” and “RandomizedSearchCV” with 10-fold cross-validation.

Since the dataset is unbalanced, I under-sampled the majority case before splitting the data into training and testing datasets. Then I used the resampled training data to train the three models. Table 1 shows the training prediction accuracy and the ROC/AUC score for each model.

Table 1 Model Evaluation

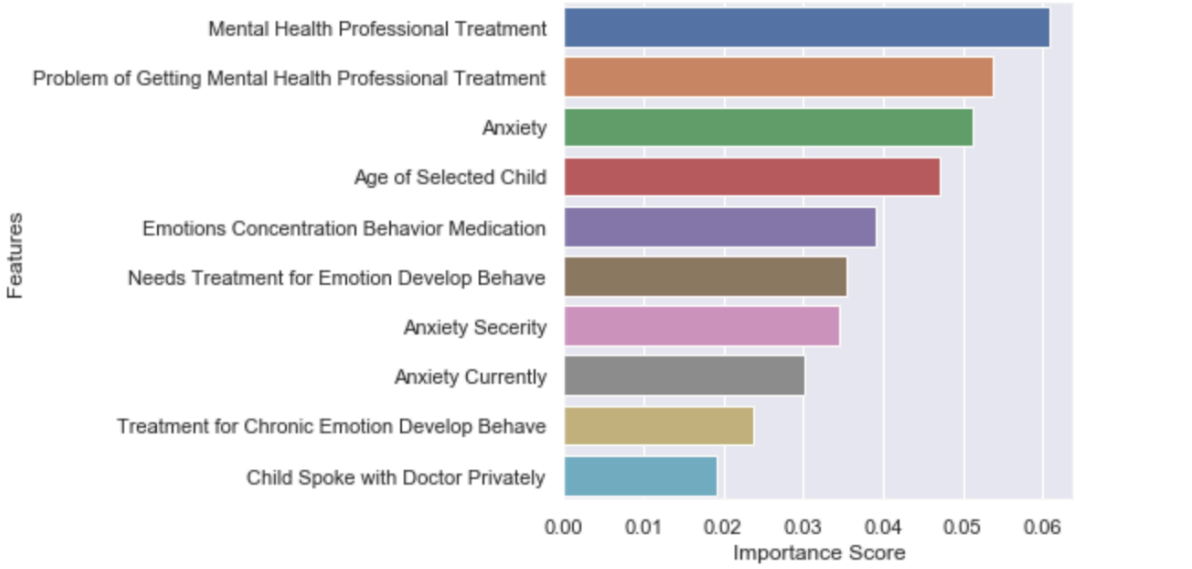
Models	Training Prediction Accuracy	Training ROC/AUC
Decision Tree	0.891	0.928
Bagging	0.770	0.861
Random Forest	0.899	0.965

From table 1 we can see that Random Forest has the highest training prediction accuracy and the ROC/AUC, and thus performs better than the other two models on the training dataset.

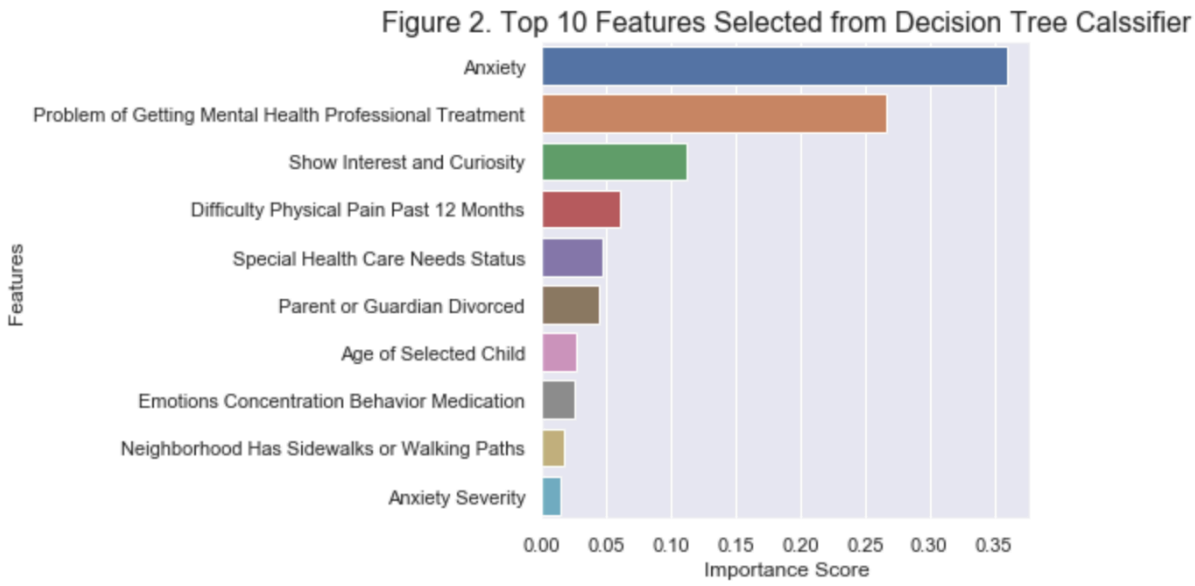
Feature Selection

To select features, I used feature importance and applied the random forest model on the training dataset. The top five features selected using random forest are mental health professional treatment, the problem of getting mental health professional treatment, anxiety diagnose, age of the selected child, and current anxiety (see Figure 1 for details). Using the selected features and random forest model on the test dataset, the minority case test accuracy reaches 93.9% and overall test accuracy is 90.1%.

Figure 1. Top 10 Features Selected from Random Forest



A decision tree classifier was also implemented to find the important features. As a result, diagnose with anxiety, mental health professional treatment, whether taken any emotions concentration behavior medication, and ADD/ADHD diagnose are considered the four most important features (see Figure 2 for details).



Decision Tree

Though the random forest model seems to perform better in the model evaluation, the tree plot using the random forest is very hard to interpret. I implemented a very simple decision tree classifier with maximum steps of 3 (see Figure 3).

Figure 3. Decision tree of identifying children at risk for depression