Predict Depressive Disorder Using CDC BRFSS Survey Data

Jingyi Lu - DSI
https://github.com/Jingyi-666/DATA1030Project

1. Introduction

Many psychological research have shown that mental disorders often trigger physical diseases. However, people are more likely to be aware of their physical diseases compared to their mental disorders. Therefore, in order to provide more timely and appropriate treatment, it is essential to solve the problem of using the available dataset to predict whether the person has any depressive disorder.

The original dataset was the information collected from the survey last year done by Behavioral Risk Factor Surveillance System (BRFSS), which is a health-related telephone survey system from CDC. The survey interviewed over 445,000 adults across the United States. Because of the size of the original dataset, a processed version from Kaggle was used, which selected 40 variables that are potentially related to heart disease over the original 300 variables. These selected variables has the potential of being associated with depressive disorders as well, but there are some variables that are redundant for constructing the models. For instance, height and weight can be cooperatively represented by BMI, while the experience of having a CT scan over chest area and vaccine records are weakly correlated with depressive disorders. Consequently, such variables should be excluded before building the machine learning pipelines.

2. EDA

The target variable is labelled as “HadDepressiveDisorder” in the dataset. The question in the original survey asked whether the person had ever diagnosed with depressive disorders with the answers of “yes”, “no”, and null values, so the variable is binary and categorical. The bar plot shows that around 20% of the participants have depressive disorders, demonstrating the imbalance of the data and thus implying that the data should be stratified while splitting.

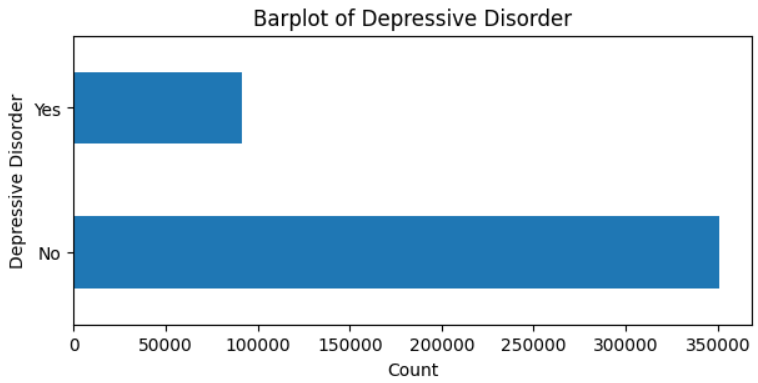

Figure 1: Bar plot of target variable with 20% of points in class 1 (“Yes”)

One critical result from the EDA is that the histograms of sleep hours for people with and without depressive disorders significantly overlap with each other, illustrating that the distributions of the sleep hours for the two groups are similar. This is surprising because people with depressive disorders tend to have sleeping issues such as oversleeping and sleep deprivation. However, the figure does show some differences for 4-5 hours of sleep, which can be considered as sleep deprivation of people with depressive disorders.

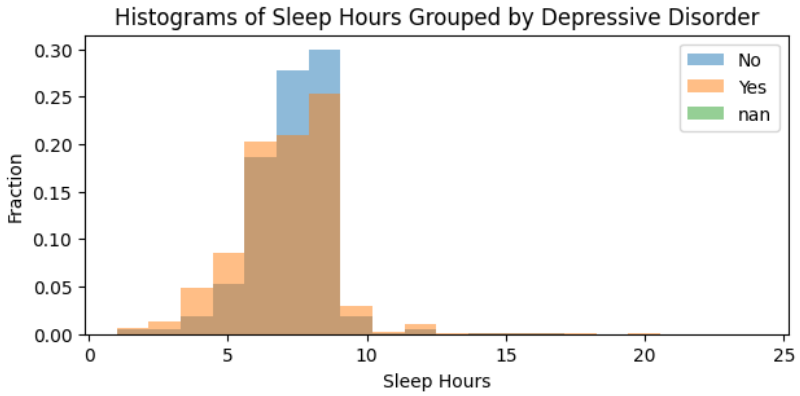

Figure 2: Histograms of sleep hours with overlapped areas for different groups of target variable

3. Methods

Analysis for missing values showed that around 35% of the data points have missing values and each feature (except State and Sex) has missing values. In consideration of the large size of the dataset and the analyzing purpose of the problem, it was appropriate to drop all rows with missing values before splitting and preprocessing. However, the dataset without null values was still very large with nearly 290,000 data points. As a result, a sample of 10% data were split out first for training machine learning models. A second split was done for separating out the 20% sample points for test set. For logistic regression, random forest classification, and nearest neighbors classification models, cross validation was utilized with folds of 4, while basic split was employed in XGBoost classification model for separating the 60% sample points for train set and 20% sample points for validation set. All of the splitting strategies mentioned above were stratified based on the target variable due to the imbalance of the dataset.

Since the feature matrix includes all of the categorical, ordinal, and numerical variables, the four transformers, ordinal encoder, one-hot encoder, min-max scaler, and standard scaler, were used for preprocessing. There were 4 numerical variables in the feature matrix and the time-related ones were preprocessed using the min-max scalar. Furthermore, a final step of standard scaler was utilized to impute all the features in a standard scale. The number of features increased from 32 to 113 after preprocessing.

The four machine learning models, Logistic Regression, Random Forest Classifier, Nearest Neighbors Classifier, and XGBoost Classifier, were chosen taking into account the large size of the dataset again. In logistic regression model, since elastic net can include both l1 and l2 regularization by changing the l1 ratio to 0 and 1, penalty was not tuned as a hyperparameter and was set with the value of “elasticnet” with a solver of “saga”. Class weight is another hyperparameter that was tuned in the model. In random forest

classification model, the maximum depth and maximum features of the tree were tuned, while in nearest neighbors classification model, the hyperparameters tuned were the number of neighbors, the weight function used in prediction, and the metric for distance computation. Finally, in the XGBoost classification model, the alpha and lambda, which are l1 and l2 regularizations, and the maximum depth of the tree were tuned. The values of these models are selected from common options.

Based on the ultimate goal of the analysis, which is to identify and provide treatment for people with depressive disorder, the logloss metric was chosen as the evaluation metric in order to capture a larger fraction of predicted positives that are true positives. Additionally, to measure the uncertainties of the evaluation metric, 5 random states were looped through to train the models

4. Results

In general, the mean baseline logloss score for all random states was 7.52 with slight variations between XGBoost classification model and other models generating from the difference in the train sets split by different strategies. The mean test scores for the four models were 0.393 (δ=0.00195), 0.384 (δ=0.00324), 0.443 (δ=0.00114), and 0.380 (δ=0.00467) respectively. Log scale of the y axis was used to illustrate the differences between mean test scores and the values were too small to be shown completely in the figure. Nevertheless, the scores showed that XGBoost classification model was the most predictive one with lowest logloss score.

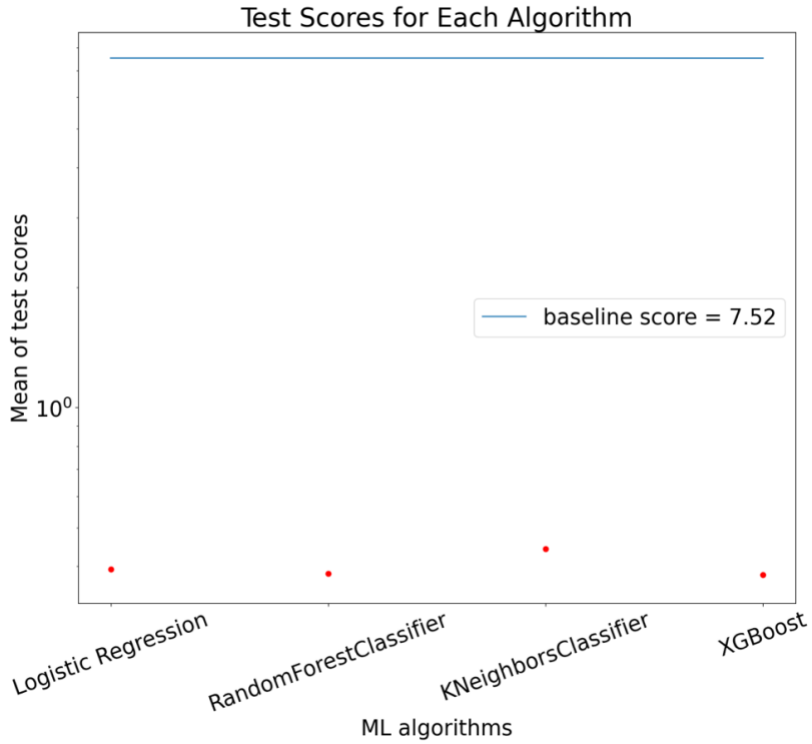

Figure 3: Logloss test scores for each model significantly lower than baseline scores; lowest score from XGBoost Classifier

The global feature importance can be demonstrated by the summary plot of SHAP values. The figure shows that MentalHealthDays, which records the number of days the mental health of the participants were self-evaluated to be not good in a 30-day period, was the

most important feature in predicting whether the person has depressive disorders or not. Besides, sex and age were also powerful features in the predictions.

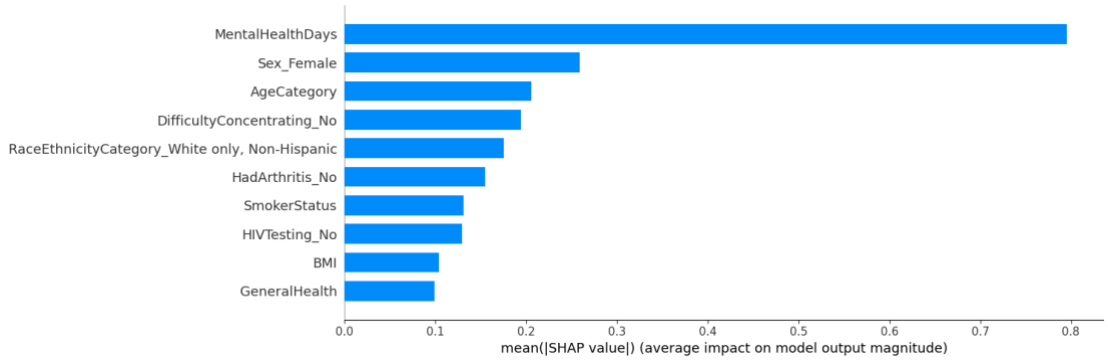

Figure 4: Global feature importance from SHAP value; MentalHealthDays as the top feature

With the most predictive model, the percentage of false positives were extremely small compared to true negatives. Although the percentage of true positives was smaller than that of false positives, the difference between the percentages were not too large compared to that between the percentages in class 0 ("No"). Therefore, the performance of the model was consistent to the expectation.

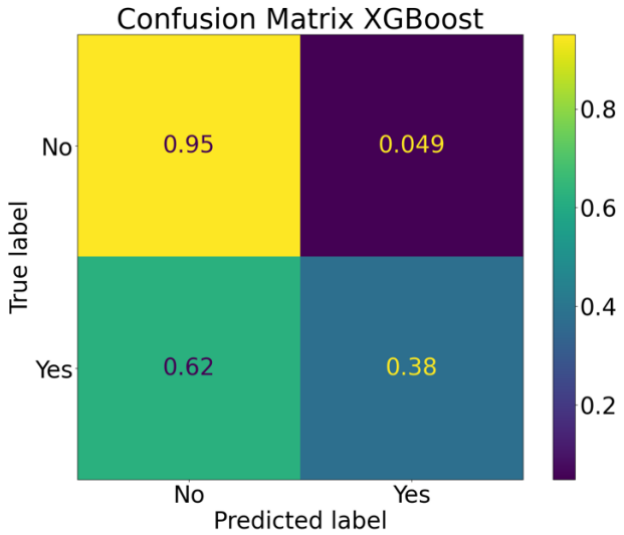

Figure 5: Confusion matrix for XGBoost classification model; very small percent of FPs; less than half of TPs

5. Outlook

Since random tree models are not ideal for large datasets, future operations can focus on other algorithms that perform better in large datasets, while computation time may be increased significantly as a result. Besides, feature importance with an emphasis on interactions between features can be employed to remove features for improving the predictive power of the models.

For details in interpretability, the specific amount of points in each category of the confusion matrix should be examined to decrease the influence of imbalance of the dataset.

6. Reference

[1]Original dataset: https://www.cdc.gov/brfss/annual_data/annual_2022.html
[2]Kaggle dataset: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data
[3]Variable documentation: https://github.com/kamilpytlak/data-science-projects/blob/main/heart-disease-prediction/2022/documentation/vars_list_with_descriptions.txt