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

Jiaxu Han

University of Chicago

Introduction

Early-onset depression is a major risk factor that surrogates multiple influences on one's developmental trajectory. It often leads to a poor academic outcome, substance abuse, suicide, and other psychiatric disorders (Birmaher et al., 1996). Therefore, understanding what might be predictive of early-onset depression is critical for early detection and developing intervention plan. It is now widely understood that singular, linear cause will rarely obtain (Cicchetti & Sroufe, 2000). The tenet of developmental psychopathology is to acknowledge the complexity of causality. Instead of trying to identify a linear causal relationship, it emphasizes on understanding the interaction of risk and protective factors on mental health outcomes. However, it is challenging for classic inferential paradigms such as group comparison or correlation study to capture the complexity of human developmental trajectories and traditional methods have raised replication and reproducibility issues (Dwyer, Falkai, & Koutsouleris, 2018). Machine learning methods, however, provides an alternative approach to analyze data with large number of variables with its own advantages such as enhanced predictability.

Research Question

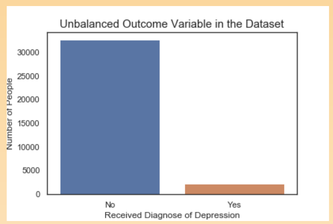
What are the risk factors for early-onset depression among children and adolescents?

Data & Methods

Data. The data used in this project is collected through the National Survey of Children's Health (NSCH). It 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 are made available to the public. In the current project, we merged 2016 and 2017 datasets.

Methods. Four resampling methods and three machine learning models are used to train the testing data. The data analysis process would follow the pipeline below:

Outcome Variable. "K2Q32A" asks whether a doctor or other health care provider has ever told the respondent that his/her child has depression. This is an unbalanced outcome variable.



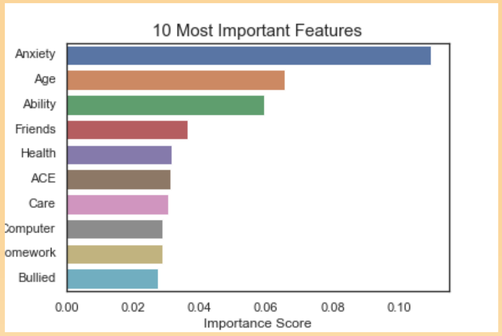
Model Selection

Random forest model coupled with a random upper sampler performs the best among all the models. It has the highest cross-validation prediction accuracy as well as the highest ROC/AUC score.

Sampler	Model	Accuracy	ROC/AUC
Random Under Sampler	Decision Tree	0.876	0.879
SMOTE	Decision Tree	0.926	0.922
ADASYN	Decision Tree	0.929	0.912
Random Over Sampler	Decision Tree	0.925	0.663
Random Under Sampler	Bagging	0.884	0.893
SMOTE	Bagging	0.880	0.912
ADASYN	Bagging	0.889	0.918
Random Over Sampler	Bagging	0.901	0.900
Random Under Sampler	Random Forest	0.866	0.943
SMOTE	Random Forest	0.949	0.943
ADASYN	Random Forest	0.949	0.943
Random Over Sampler	**Random Forest**	**0.950**	**0.944**

Feature Selection

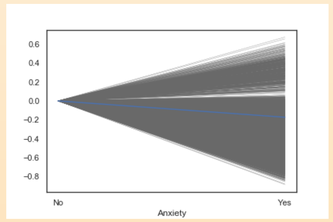
We selected 10 important features using Random Forest on the whole dataset.

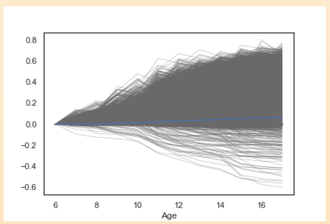


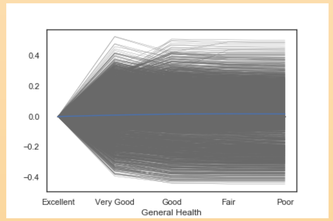
Global Interpretation

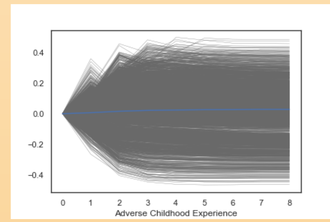
We use Individual Conditional Expectation (ICE) plots and Partial Dependence Plots (PDP) to visualize the dependence of the prediction on each one of the ten selected features.

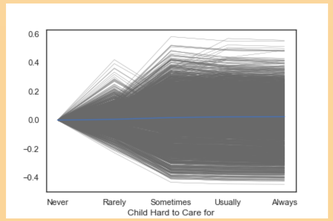
ICE and PDP Plots for the Selected 10 Features

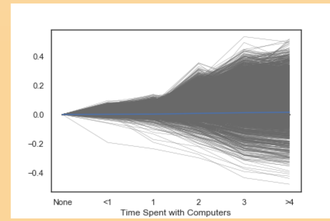


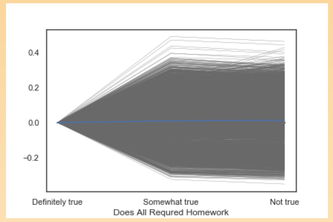


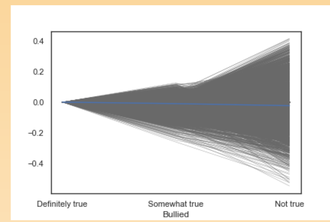






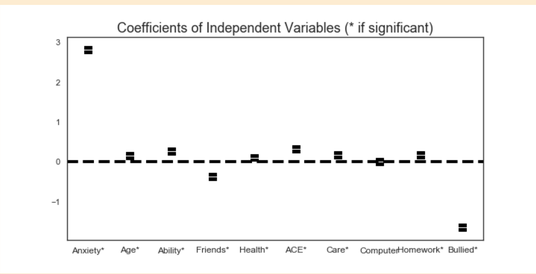






Logistic Regression

A logistic regression was implemented on the whole dataset using the ten features as independent variables



Conclusions

- A diagnose of anxiety is significantly predictive of having a diagnose of depression.
- There is a higher probability of getting a diagnose of depression as the child gets older.
- Poor health, difficulty of making friends, being bullied at school as well as more adverse childhood experience are all predictive of depression.
- When a parent reports that the child is hard to care for or not doing homework, it may suggest that the child have depression.
- Though time spent with computers was selected as an important feature, it is not significantly predictive of depression.

References

- Birmaher B, Ryan N.D., Williamson D.E., Brent D.A., Kaufman J, Dahl R.E., Perel J, Nelson B. (1996). Childhood and Adolescent Depression: A Review of the Past 10 Years. Part I. *Journal of the American Academy of Child & Adolescent Psychiatry*, *35*(11), 1427–1439.
- Cicchetti, D., & Sroufe, L. A. (2000). The past as prologue to the future: the times, they've been a-changin'. *Development and Psychopathology*, *12*(3), 255–264.
- Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118.
- US Census Bureau. (2019, April 23). National Survey of Children's Health (NSCH). Retrieved from https://www.census.gov/programs-surveys/nsch.html