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

Final Paper for MACS 30200 Perspectives on Computational Research

Jiaxu Han

Abstract

Early-onset depression is a common and potentially debilitating disorder for children and youth. The present paper uses a machine learning model to identify risk factors for the disorder. The dataset is from the National Survey of Children's Health (NSCH). After data preprocessing, a random forest model was applied on the whole cleaned dataset to conduct feature selection. Then, a global interpretation and logistic regression are applied using the top 10 important features. Results show that a diagnose of anxiety is significantly predictive of having a diagnosis of depression. There is a higher probability of getting a diagnose of depression as the child gets older. Poor health, the difficulty of making friends, being bullied at school are all predictive of depression. A single childhood adverse event seems not powerful enough to predict early-onset depression, but a combination of multiple adverse events can. 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.

Introduction

According to a national comorbidity survey (Kessler et al., 2005), about half of the population in the U.S. will meet the criteria for a DSM-IV diagnosis sometime in their life. The survey also reports that the most prevalent mental illness is a major depressive disorder with a

16.6% lifetime prevalence rate. The first onset of a depressive episode often occurs in childhood or adolescence. A series of research studies have shown that early-onset depression is not only a result of the cumulative effect of various kinds of stressors in life but also serves as a major risk factor that surrogates multiple influences on one's developmental trajectory. It is very likely for children with depression to develop more severe depression and other comorbid psychiatric disorders in adulthood (Birmaher et al., 1996). Besides, it is predictive of chronic physical conditions (Scott et al., 2011), drug abuse (Reinherz et al., 2000), poor academic outcome, suicide (Birmaher et al., 1996), future negative life events (Patton et al., 2003) and so on. Given the prevalence and the potential negative impact of depression, there has been a heightened awareness among parents, teachers, as well as mental health professionals. Identifying risk factors and symptoms that precedes a full-blown depressive episode, therefore, becomes critical for developing intervention and treatment plan.

Many research studies choose to use classic inferential paradigms such as group comparison or correlation study to examine the relationship between a proposed risk factor and depression. Many agree that prevention is better than cure. However, developing comprehensive preventive strategies for early-onset depression only using classical inferential approach has been challenging. First, it is almost impossible to find an exact cause for depression. It is now widely understood that singular, linear cause will rarely obtain (Cicchetti & Sroufe, 2000). In addition, unlike many physical diseases, there is no biomarker that has been identified for the diagnosis of depression. Therefore, using classical inferential paradigms is not ideal for capturing the complexity of human developmental trajectories and often raise application and reproducibility issues (Dwyer, Falkai, & Koutsouleris, 2018).

An alternative approach to classical inferential methods would be machine learning. In the past, due to limited resource, a single research study can only examine so many variables. The classic inferential approach is well suited for this kind of research, but the results do not apply to each individual. In recent years, there has been increased collaboration between multiple sites across the country. With collaborative efforts, the research study is now able to gather much more information about each participant and build a large dataset with thousands of participants. Machine learning has its own advantages such as enhanced predictability in analyzing a dataset that contains hundreds of or even thousands of variables.

Using machine learning models, the current paper is trying to answer two questions: (1) what are the risk factors that are most predictive of early-onset depression among children and adolescents? (2) how does each risk factor associated with the possibility of having depression? The objective of the current study is to provide parents, teachers, and mental health professionals a short list of risk factors to quickly identify children or youth at risk for depression and provide support and intervention before things deteriorate.

Data

The project uses a dataset that 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 only uses 2016 and 2017 dataset 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

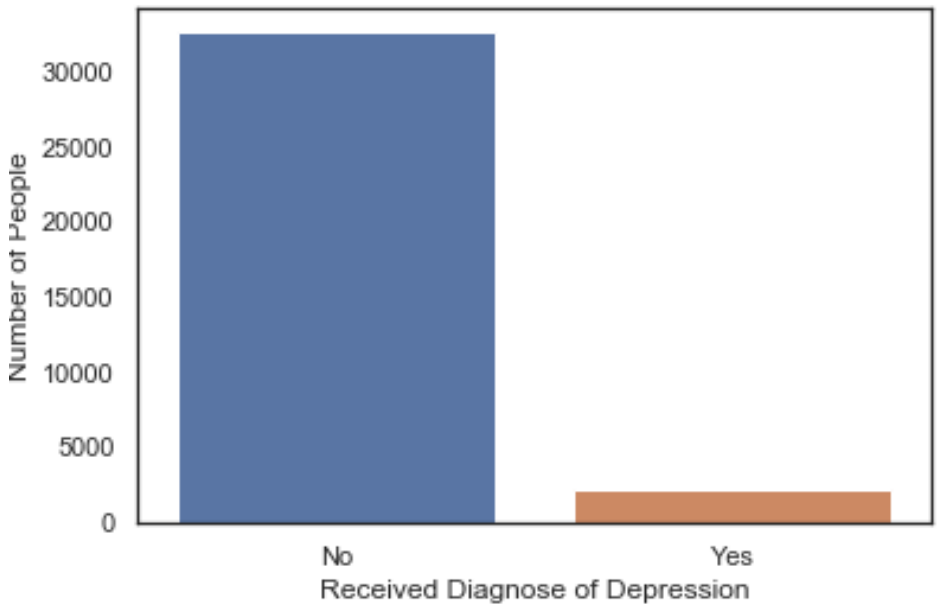
There are three questions associated with the child's depression in the survey: 'K2Q32A' asks whether a doctor or other health care provider has ever told the respondent that his/her child has depression; 'K2Q32B' asks the respondent whether his/her child currently has depression; and 'K2Q32C' asks the severity of the depression. Here I chose whether or not a child has been diagnosed with depression as the main outcome variable. Again, 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, I took a more stringent approach to preprocess the dataset. I first manually selected 90 variables from 429 variables in the main dataset. Those variables have been considered relevant to early-onset depression based on prior literature review. The selected variables include some questions about general and certain health condition, psychological and temperamental factors, family and school

environment, and several demographic questions, and so on. There are still many missing values in responses to the selected variables. Again, considering the unbalanced nature of the dataset and most of the variables are discrete variables with only two choices, imputing missing values using either mean/median or the most frequent values will result in bias. Therefore, I decided to simply drop columns with the response rate lower than 50% and then I dropped rows with any missing values. Finally, I added an additional variable, cumulative adverse childhood experience (ACE), by adding up 8 discrete ACE variables ('ACE3', 'ACE4', 'ACE5', 'ACE6', 'ACE7', 'ACE8', 'ACE9', 'ACE10') from the dataset. Those discrete ACE variables were also included in the final analysis. As a result, a sample size of 34821 and 78 variables are included for data analysis. Now, the majority of cases versus minority cases ratio drops to around 15:1 (see Figure 1).

FIGURE 1. THE FINAL DATASET HAS AN UNBALANCED OUTCOME VARIABLE



Descriptive Statistics

In the cleaned dataset, the mean age of the children is 12.10, ranging from 6 to 17. 79.35% of the children are white, 5.49% are black, and 4.77% are Asian. Only 6.17% of the respondents reported a child having depression.

Models

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 imbalanced 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. I implemented those three models using the *sklearn* package in Python3.

In addition, to counter the imbalanced nature of the dataset, I tried four different samplers coupled with above three machine learning models on the training dataset. All four samplers are implemented using *imblearn* package. The four samplers are RandomUnderSampler, RandomOverSampler, SMOTE, and ADASYN. RandomUnderSampler is an under-sampling method which under-sample the majority class by randomly picking samples with or without replacement. Similarly, RandomOverSampler over-sample the minority class by picking samples at random with replacement. SMOTE is a synthetic minority over-sampling technique presented by Chawla and colleagues (2002). ADASYN also performs over-sampling but using Adaptive Synthetic sampling approach developed by He and colleagues (2008). Methods

Hyper-parameter Tuning and Model Evaluation

Since there are four kinds of samplers and three machine learning models, I evaluated twelve models in total. Before training the models, I used the *sklearn* package built-in method *RandomizedSearchCV* with 10-fold cross-validation to find the suitable parameter grid based training prediction accuracy for each of the models. In the Appendix, it shows the optimal parameter grid searched by *RandomizedSearchCV.*

Then, I trained each model using the parameter grid in Table A1. The training prediction accuracy and the ROC/AUC score for each model is presented in Table 1. From Table 1 we can see that Random Forest paired with *RandomOverSampler* has the highest training prediction accuracy and the ROC/AUC score. Thus, it performs better than all the other models on the training dataset. I care more about the ROC/AUC score here because it tells how much the model is capable of distinguishing between classes and the dataset is imbalanced. It seems that the random forest model with *RandomOverSampler* has a satisfying ROC/AUC score of 0.94. In sum, I decided to use *RandomOverSampler* and Random Forest model to do feature selection in the following.

TABLE 1. TRAINING ACCURACY AND ROC/AUC USING 10-FOLD CROSS-VALIDATION

Sampler	Model	Prediction Accuracy	ROC/AUC
RandomUnderSampler	Decision Tree	0.876	0.879
SMOTE	Decision Tree	0.926	0.922
ADASYN	Decision Tree	0.929	0.912
RandomOverSampler	Decision Tree	0.925	0.663
RandomUndeSampler	Bagging	0.884	0.893

SMOTE	Bagging	0.880	0.912
ADASYN	Bagging	0.889	0.918
RandomOverSampler	Bagging	0.901	0.900
RandomUnderSampler	Random Forest	0.866	0.943
SMOTE	Random Forest	0.949	0.943
ADASYN	Random Forest	0.949	0.943
RandomOverSampler	Random Forest	0.950	0.944

Feature Selection

As mentioned above, the cleaned dataset has a total of 78 variables. Though it is possible that each one of the 77 predictive variable is associated with the outcome variable, it is very redundant and contrary to the objective of this study. Therefore, I applied random forest model and feature importance on the resampled whole dataset to select the most important features from the dataset.

The concept of feature importance is very straightforward: the importance of a feature is the increase in the prediction error of the model after permuting the feature (Molnar, 2019). A feature is important if the prediction error is unchanged after shuffling its values (Molnar, 2019). In this study, I used a built-in method in *sklearn* package to select top 10 important features from the 77 predictive variables.

Relationship between Top 10 Important Features and Outcome Variable

As mentioned above, the cleaned dataset has a total of 78 variables. Though it is possible that each one of the 77 predictive variables is associated with the outcome variable, it is very redundant and contrary to the objective of this study. Therefore, I applied the random forest

model and feature importance on the resampled whole dataset to select the most important features from the dataset.

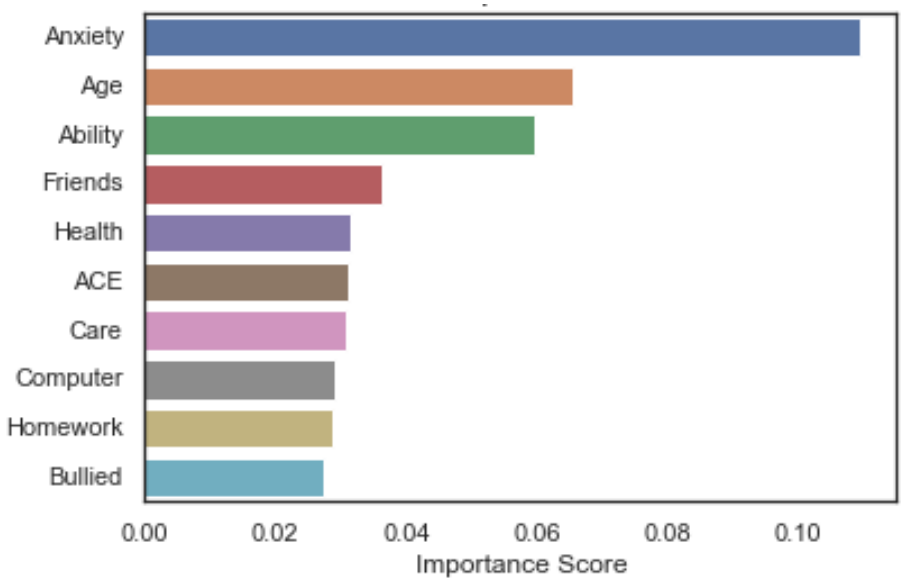
The concept of feature importance is very straightforward: the importance of a feature is the increase in the prediction error of the model after permuting the feature (Molnar, 2019). A feature is important if the prediction error is unchanged after shuffling its values (Molnar, 2019). In this study, I used a built-in method in *sklearn* package to select the top 10 important features from the 77 predictive variables.

Results

Feature Importance

According to the result from feature importance (see Figure 2), the top 10 important features are *anxiety*, *age of the child*, *health affected ability*, *difficulty making friends*, *general health*, *cumulative adverse childhood experience*, *child hard to care for*, *time spent with computers*, *does all required homework*, and *bullied*.

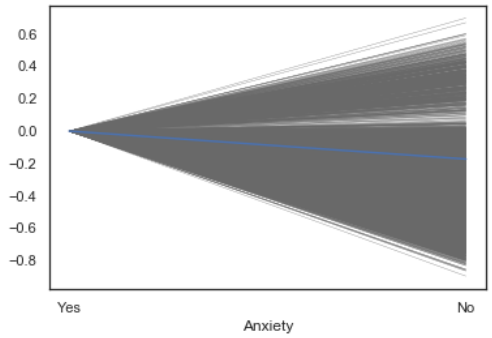
FIGURE 2. TOP 10 IMPORTANT FEATURES

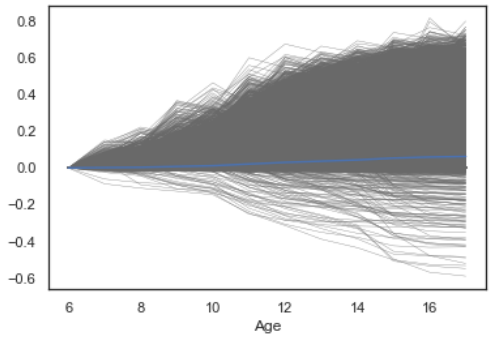


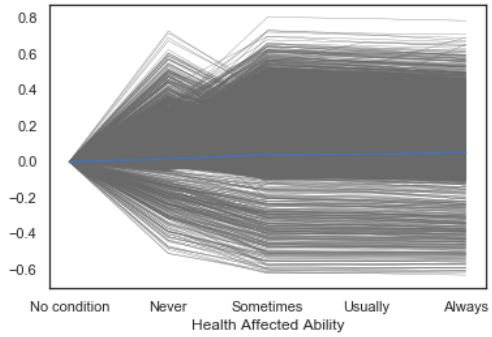
PDP/ICE plots

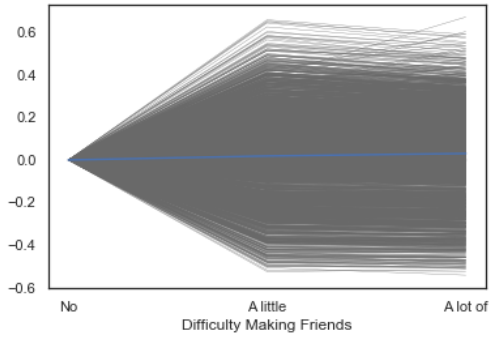
Figure 3 presents the PDP and centered ICE plots for each one of the selected features. From the plots we can see that: (1) overall, children who do not have anxiety are less likely to have depression, and vice versa; (2) the probability of having depression increases as children's age increases; (3) overall, there is a slightly increased probability of having depression as the degree of the health condition that affects ability gets larger; (4) overall, there is only a slight increase in the probability of having depression if children have an increased level of difficulty making friends; (5) overall, there is a slight increase in the probability of having depression as the children's health condition deteriorates; (6) As the adverse childhood experience accumulates, the probability of having depression also slightly increase; (7) As the frequency of parents report that the child is hard to care for increases, the probability of the child has depression also increases; (8) the PDP plot is almost horizontal to the x-axis for feature "time spent with computers". The ICE instances suggest that more children have an increased likelihood of having depression when they spent more time with computers; (9) it is very hard to tell how 'does all required homework' affect the probability of having depression from the PDP/ ICE plot, it may require further analysis to understand the relationship between them; (10) it is also not very clear how being bullied at school affects probability of having depression from the PDP/ICE plot. Further analysis is needed.

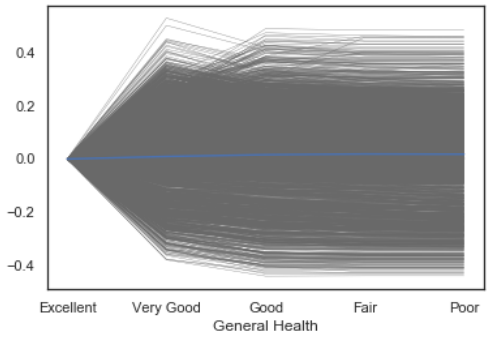
FIGURE 3. PDP/ICE PLOTS FOR 10 IMPORTANT FEATURES

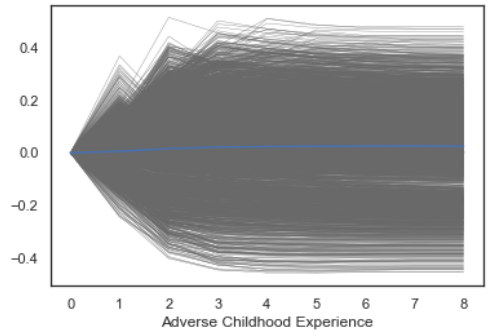

0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Yes
No
Anxiety

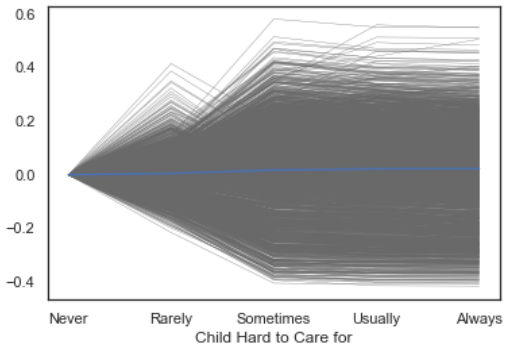

0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
6
8
10
12
14
16
Age

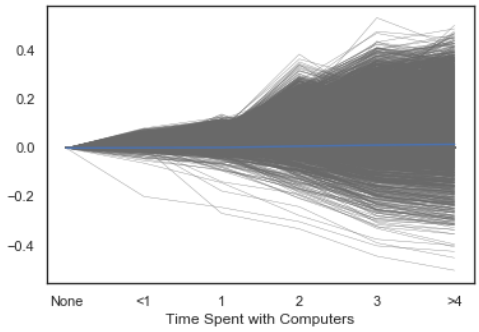

0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
No condition
Never
Sometimes
Usually
Always
Health Affected Ability

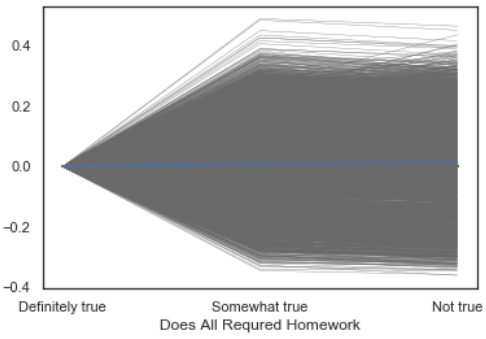

0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
No
A little
A lot of
Difficulty Making Friends

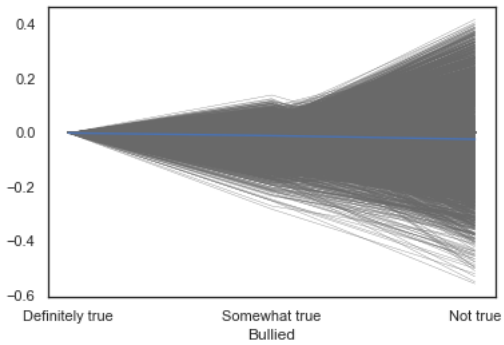

0.4
0.2
0.0
−0.2
−0.4
Excellent
Very Good
Good
Fair
Poor
General Health


0.4
0.2
0.0
−0.2
−0.4
0
1
2
3
4
5
6
7
8
Adverse Childhood Experience


0.6
0.4
0.2
0.0
−0.2
−0.4
Never
Rarely
Sometimes
Usually
Always
Child Hard to Care for


0.4
0.2
0.0
−0.2
−0.4
None
<1
1
2
3
>4
Time Spent with Computers

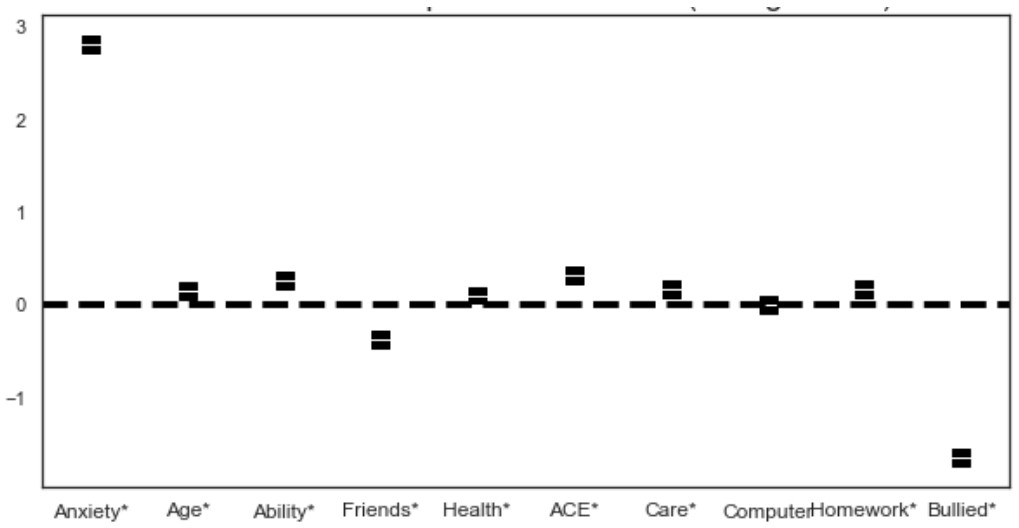




Logistic Regression

The result of logistic regression (see Figure 4) shows that except for *time spent with computers*, all the other 9 variables are significantly correlated with the probability of having a diagnosis of depression. From Figure 4 we can see that (1) having anxiety is positively correlated with the probability of having depression ($r = 2.81$, $p < .01$); (2) children's age is positively correlated with the probability of having depression ($r = 0.15$, $p < .01$; (3) the degree of health condition that affects ability is positively correlated with the probability of having depression ($r = 0.26$, $p < .01$); (4) having less difficulty of making friends is negatively correlated with probability of having depression ($r = -0.36$, $p < .01$); (5) as health condition deteriorates, the probability of having depression increases ($r = 0.10$, $p < .01$); (6) cumulative adverse childhood experience is significantly correlated with increased probability of having depression ($r = 0.33$, $p <. 01$); (7) 'Child hard to care for' is also positively correlated with the probability of having depression ($r = 0.18$, $p < .01$); (8) 'Time spent with computers' is not significantly predictive of the probability of having depression; (9) 'Does all required homework' is positively correlated with probability of having depression ($r = 0.17$, $p = 0.14$) and (10) being bullied at school is also predictive of having depression ($r = -1.64$, $p < .01$).

FIGURE 4. COEFFICIENTS OF INDEPENDENT VARIABLES (* IF SIGNIFICANT)



Post hoc Analysis

To examine whether those risk factors identified above are specific for depression or not, I conducted an additional analysis using anxiety as the outcome variable. I used the same pipeline as discussed above to identify risk factors for anxiety: hyper-parameter tuning, model selection (the same four samplers and three machine learning models), and feature importance. The analysis revealed that the top 10 important features for anxiety are: *health affects ability*, *depression*, *stays calm and in control when challenged*, *difficulty making friends*, *child hard to care for*, *child's age*, *general health*, *bullied*, *difficulty concentrating, remembering, or making decisions* and *ADD/ADHD.*

Conclusion

Among the top ten risk factors identified using random forest model, nine of them are significantly predictive of the probability of having depression. Those significant risk factors for depression are having anxiety, children's age, poor health condition, ability significantly affected by health, difficulty making friends, cumulative adverse childhood experience, child hard to care for, difficulty finishing homework, and being bullied at school. To examine if those risk factors are specific for predicting depression, I trained another random forest model to identify risk factors for anxiety — another mood disorder that is often comorbid with depression. The result of post hoc analysis suggests that some risk factors may be shared by mood disorders in general, such as age, health affects ability, difficulty making friends, child hard to care for, and bullied at school. Some risk factors may be more specific for predicting depression: general health, cumulative adverse childhood experience, and does all homework. Anxiety is a predictor for depression and vice versa.

Discussion

Anxiety and depression are both mood disorders and depression have been found to be highly comorbid with anxiety in both clinical practice and research. Therefore, it is not surprising that they seem to be predictive of each other in this dataset. Given such information, it is important to pay attention to early signs of any type of mood disorder in children. The study also further explores the risk factors for both mood disorders, although it is mainly focused on depression. Depression and anxiety apparently share some common risk factors. Age is a major risk factor. As children grow older, there is an increased possibility of having a diagnose of mood disorder. Age probably does not cause mood disorders directly. It is more likely that children would encounter more and more stressors as they grow up, and without appropriate intervention,

there might be some cumulative effects of those stressors that are manifested in one's life as depression or anxiety. Health affected ability is also a general risk factor for a mood disorder. Without further information, it is hard to conclude how does this factor contribute to the development of mood disorders. Except for rare situations where certain diseases or medication can actually cause depression or anxiety symptoms, it is likely that children whose ability are affected by poor health condition would miss out the kind of life and childhood that other kids are living. Not being able to go to school, being laughed at by other kids, difficulty making friends, and so on are all possible repercussions of poor health condition and lead to depression or anxiety.

Difficulty making friends and being bullied at school reflects the social components of the risk factors. Many research has also confirmed the importance of friendship on promoting healthy development for children. Parents and teachers, therefore, may want to pay more attention to children's social experience at school and create a safe environment for children.

It is interesting to see that "child hard to care for" is a risk factor for mood disorders. This is the only risk factor that reflects a parent's feeling for the child. Feeling that a child is hard to care for can go both ways for the development of mood disorders. First, if a child has already developed depression or anxiety symptoms, it might lead the parent to feel that his/her child need more attention to care for. Second, it is equally possible that the parent's feeling might be a risk factor. Children may perceive the parents' frustration and stress as a stressor which can lead to depressive or anxiety symptoms. It is advised that mental health professionals explore the parents' feeling about the child and where does that come from.

Health condition, cumulative adverse childhood experience, and doing homework seem to be more predictive of depression than that of anxiety symptoms. It all makes sense that if a child has poor health, had much negative childhood experience, and often couldn't finish his/her homework, it may suggest that he/she has depressive symptoms. Note that cumulative adverse childhood experience instead of a single one that is predictive of depression. This is consistent with many findings in the developmental psychopathology. One single negative childhood experience may not ruin one's life if given enough support and resources. However, if a child has multiple stressors in life and could not get the needed help, it might progress into a larger problem later.

References

Abrams, R. C., Rosendahl, E., Card, C., & Alexopoulos, G. S. (1994). Personality Disorder Correlates of Late and Early Onset Depression. *Journal of the American Geriatrics Society*, *42*(7), 727–731.

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.

Côté, S. M., Boivin, M., Liu, X., Nagin, D. S., Zoccolillo, M., & Tremblay, R. E. (2009). Depression and anxiety symptoms: Onset, developmental course and risk factors during early childhood. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *50*(10), 1201–1208.

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.

Gollan, J., Raffety, B., Gortner, E., & Dobson, K. (2005). Course profiles of early- and adult-onset depression. *Journal of Affective Disorders*, *86*(1), 81–86.

Hanna, E. Z., Yi, H. Y., Dufour, M. C., & Whitmore, C. C. (2001). The relationship of early-onset regular smoking to alcohol use, depression, illicit drug use, and other risky behaviors during early adolescence: Results from the youth supplement to the Third National Health and Nutrition Examination Survey. *Journal of Substance Abuse*, *13*(3), 265–282.

He, Haibo, Yang Bai, Edwardo A. Garcia, and Shutao Li. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," In IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322-1328, 2008.

Jane Costello, E., Erkanli, A., & Angold, A. (2006). Is there an epidemic of child or adolescent depression? *Journal of Child Psychology and Psychiatry*, *0*(0), 1263-71.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. *Archives of General Psychiatry*.

Korczak, D. J., & Goldstein, B. I. (2009). Childhood Onset Major Depressive Disorder: Course of Illness and Psychiatric Comorbidity in a Community Sample. *Journal of Pediatrics*, *155*(1), 118–123.

Korten, N. C. M., Comijs, H. C., Lamers, F., & Penninx, B. W. J. H. (2012). Early and late onset depression in young and middle aged adults : Differential symptomatology , characteristics and risk factors ? *Journal of Affective Disorders*, *138*(3), 259–267.

Levinson, D. F., Zubenko, G. S., Crowe, R. R., DePaulo, R. J., Scheftner, W. S., Weissman, M. M., ... Chellis, J. (2003). Genetics of recurrent early-onset depression (GenRED): Design and preliminary clinical characteristics of a repository sample for genetic linkage studies. *American Journal of Medical Genetics*, *119B*(1), 118–130.

Lewinsohn, P. M., Rohde, P., & Seeley, J. R. (1998). Major depressive disorder in older adolescents: Prevalence, risk factors, and clinical implications. *Clinical Psychology Review*, *18*(7), 765–794.

Lizardi, H., Klein, D. N., Ouimette, P. C., Riso, L. P., Anderson, R. L., & Donaldson, S. K. (1995). Reports of the Childhood Home Environment in Early-Onset Dysthymia and Episodic Major Depression. *Journal of Abnormal Psychology*, *104*(1), 132–139.

Luby, J. L., Belden, A. C., & Spitznagel, E. (2006). Risk factors for preschool depression: The mediating role of early stressful life events. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *47*(12), 1292–1298.

Moffitt, T. E., Caspi, A., Harrington, H., Milne, B. J., Melchior, M., Goldberg, D., & Poulton, R. (2007). Generalized anxiety disorder and depression: Childhood risk factors in a birth cohort followed to age 32. *Psychological Medicine*, *37*(3), 441–452.

Molnar, C. (2019, April 12). Interpretable Machine Learning. Retrieved from https://christophm.github.io/interpretable-ml-book/

N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, 321-357, 2002.

Patton, G. C., Coffey, C., Posterino, M., Carlin, J. B., & Bowes, G. (2003). Life events and early onset depression: Cause or consequence? *Psychological Medicine*, *33*(7), 1203–1210.

Reinherz, H. Z., Giaconia, R. M., Carmola Hauf, A. M., Wasserman, M. S., & Paradis, A. D. (2000). General and specific childhood risk factors for depression and drug disorders by early adulthood. *Journal of the American Academy of Child and Adolescent Psychiatry*, *39*(2), 223–231.

Roane, B. M., & Taylor, D. J. (2017). Adolescent Insomnia as a Risk Factor for Early Adult

Depression and Substance Abuse. *Sleep*.

Scott, K. M., Von Korff, M., Angermeyer, M. C., Benjet, C., Bruffaerts, R., de Girolamo, G., ...

Kessler, R. C. (2011). Association of childhood adversities and early-onset mental disorders with adult-onset chronic physical conditions. *Archives of General Psychiatry*, *68*(8), 838– 844.

Smith, D. J., Duffy, L., Stewart, M. E., Muir, W. J., & Blackwood, D. H. R. (2005). High harm avoidance and low self-directedness in euthymic young adults with recurrent, early-onset depression. *Journal of Affective Disorders*, *87*(1), 83–89.

Thapar, A., Collishaw, S., Pine, D. S., & Thapar, A. K. (2012). Depression in adolescence. *The Lancet*, *379*(9820), 1056–1067.

US Census Bureau. (2019, April 23). National Survey of Children's Health (NSCH). Retrieved from https://www.census.gov/programs-surveys/nsch.html

Weissman, M. M., Gammon, G. D., John, K., Merikangas, K. R., Warner, V., Prusoff, B. A., & Sholomskas, D. (1987). Children of Depressed Parents. *Archives of General Psychiatry*, *44*(10), 847.

Weissman, M. M., Warner, V., Wickramaratne, P., & Prusoff, B. A. (1988). Early-onset major depression in parents and their children. *Journal of Affective Disorders*, *15*(3), 269–277.

Yap, M. B. H., Pilkington, P. D., Ryan, S. M., & Jorm, A. F. (2014). Parental factors associated with depression and anxiety in young people: A systematic review and meta-analysis. *Journal of Affective Disorders*. Elsevier.

Zisook, S., John Rush, A., Albala, A., Alpert, J., Balasubramani, G. K., Fava, M., ... Wisniewski, S. (2004). Factors that differentiate early vs. later onset of major depression disorder. *Psychiatry Research*, *129*(2), 127–140.

Appendix

TABLE A1. PARAMETER GRID SEARCHED BY 'RANDOMSEARCHCV'

Sampler	Model	Parameters
RandomUnderSampler	Decision Tree	'splitter': 'best' 'max_depth': 2 'criterion': 'gini'
SMOTE	Decision Tree	'splitter': 'best' 'max_depth': 7 'criterion': 'entropy'
ADASYN	Decision Tree	'splitter': 'best' 'max_depth': 7 'criterion': 'gini'
RandomOverSampler	Decision Tree	'splitter': 'random' 'max_depth': None 'criterion': 'gini'
RandomUnderSampler	Bagging	'n_estimators': 100 'max_samples': 5 'max_features': 0.2 'bootstrap_features': False 'bootstrap': False
SMOTE	Bagging	'n_estimators': 100 'max_samples': 5 'max_features': 0.8 'bootstrap_features': True 'bootstrap': True
ADASYN	Bagging	'n_estimators': 100 'max_samples': 6 'max_features': 1.0 'bootstrap_features': False 'bootstrap': False
RandomOverSampler	Bagging	'n_estimators': 200 'max_samples': 4 'max_features': 0.6 'bootstrap_features': False 'bootstrap': True

RandomUnderSampler	Random Forest	'n_estimators': 1600 'min_samples_split': 2 'min_samples_leaf': 1 'max_features': 'auto' 'max_depth': 10 'bootstrap': True
SMOTE	Random Forest	'n_estimators': 600 'min_samples_split': 5 'min_samples_leaf': 4 'max_features': 'sqrt' 'max_depth': 20 'bootstrap': False
ADASYN	Random Forest	'n_estimators': 1200 'min_samples_split': 5 'min_samples_leaf': 4 'max_features': 'sqrt' 'max_depth': 30 'bootstrap': False
RandomOverSampler	Random Forest	'n_estimators': 800 'min_samples_split': 10 'min_samples_leaf': 1 'max_features': 'auto' 'max_depth': 50 'bootstrap': True