Behavioral and Conduct Problems in Children

Holly Erickson

DSC 680 Applied Data Science

GitHub Portfolio URL: https://github.com/Holly-E/Conduct_Behavioral_Problems_Study

Milestone 2

Domain
Any surprises from your domain from these data?

The domain is conduct and behavioral problems in children under 18 years old (Target K2Q34A in the survey).

The biggest surprise from this domain has been that spending time thinking about it makes you care about it more. It makes you more curious to look at what are the environmental factors impacting these kids and what could be done to move them out of the group of behavioral and conduct problems.

For instance, one of the main predictors of K2Q34A across 2016 - 2018 is a high difficulty making friends compared to children their age. I think making friends is a skill that people take for granted, but perhaps in young children more time could be spent on teaching skills such as compassion and sharing, and encouraging activities where children can bond and make friends.

It's a really interesting feeling when your data starts to feel less like data and more like people. This is part two of a project on this survey on children, the first part being on masking in ADHD. I would love to have a career where the data is as interesting as these two projects have been.

Data
The dataset is what you thought it was?

The datasets for the years 2016, 2017 and 2018 were smaller than I expected them to be, based off of the 2012 dataset that I had used for an earlier project. They made some major changes between 2012 and 2016 and further reading explained the change in size.

The NSCH was redesigned by HRSA's Maternal and Child Health Bureau in 2016. It is now an annual survey that is fielded by the U.S. Census Bureau. A screener questionnaire is now completed, then eligible children are selected for the topical questionnaire, which has the questions I am most interested in. (In prior years all children received the full set of questions.)

For example, in 2016 a total of 138,009 screener questionnaires were completed, and 67,047 of those were eligible for topical questionnaire follow-up. Of those topical-eligible households, 50,212 completed the topical questionnaire. Children with special health care needs and young children (0 to 5 years old) had a higher probability of selection for the topical survey.

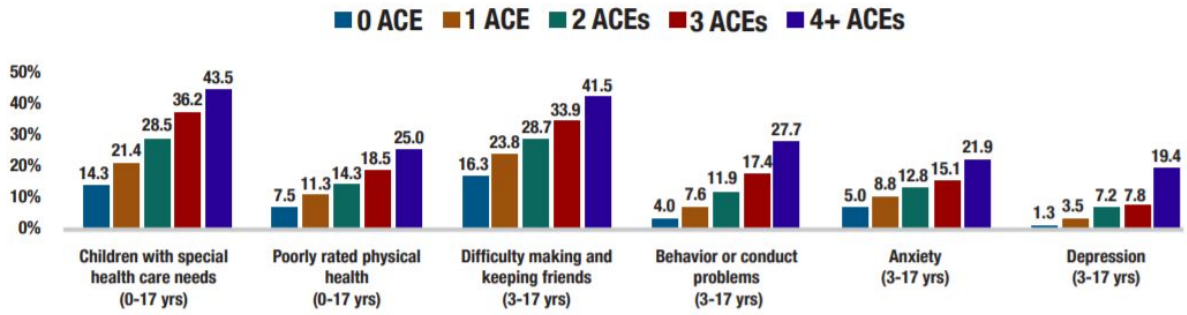
Research Questions? Benefits? Why analyze these data?
Have you had to adjust your approach or research questions?

One interesting topic I found when researching my project is the idea of the Adverse Childhood Experience (ACE) impacting K2Q34A. ACEs are stressful events that can result in chronic stress without mitigating or buffering support. This prolonged activation of the body's stress response system can negatively impact short- and long-term health and well-being.

These are the ACEs examined in the survey:

- Parent/guardian divorced or separated
- Parent/guardian died
- Parent/guardian served time in jail
- Saw or heard parents or adults slap, hit, kick, punch one another in the home
- Was a victim of violence or witnessed violence in his or her neighborhood
- Lived with anyone who was mentally ill, suicidal, or severely depressed
- Lived with anyone who had a problem with alcohol or drugs
- Was treated or judged unfairly because of his or her race or ethnic group

It's clear there is a correlation between the number of ACEs a child experiences and Behavioral and Conduct Problems:



I am deepening my research questions to examine which ACE's or combination of ACE's have the highest impact on K2Q34A.

Reference:

HRSA Maternal and Child Health. (June 2020). *"Adverse Childhood Experiences NSCH Data Brief."* Retrieved from https://mchb.hrsa.gov/sites/default/files/mchb/Data/NSCH/nsch-ace-databrief.pdf

Method
Is your method working?

The idea behind the method is working. Interestingly, the changes to the survey over time are not making models more accurate over time. They have added helpful features such as age of child when they stopped breastfeeding, but have removed helpful features such as mother's age.

As a result, my model for 2012 achieved the highest accuracy. This may also be due to changes in the way children were selected for topical questions. For all four years analyzed, models predicting K2Q34A in females are more accurate than for males.

Originally, I wanted to prepare a model that is an ensemble voting classifier that combines a Random Forest Classifier, Logistic Regression and Support Vector Machines to predict K2Q34A. However I am finding SVM's are underperforming compared to the other two types of algorithms across all years. I will try fine-tuning the SVM algorithm with hyper parameter tuning, and a smaller input of features. If that fails I will leave out the SVM algorithm.

Challenges
What challenges are you having?

The biggest challenge I am having is to be careful of changes in the dataset questions and changes in what feature values mean. For example, in 2016 and 2017 the answers to the question on whether a child is bullied are:
1 = Definitely true
2 = Somewhat true
3 = Not true

In 2018 the questions on whether a child is bullied can have the following responses:
1 = Never (in the past 12 months)
2 = 1-2 times (in the past 12 months)
3 = 1-2 times per month
4 = 1-2 times per week
5 = Almost every day

As a result, a positive correlation in 2016 and 2017 will have a similar meaning to a negative correlation in 2018.