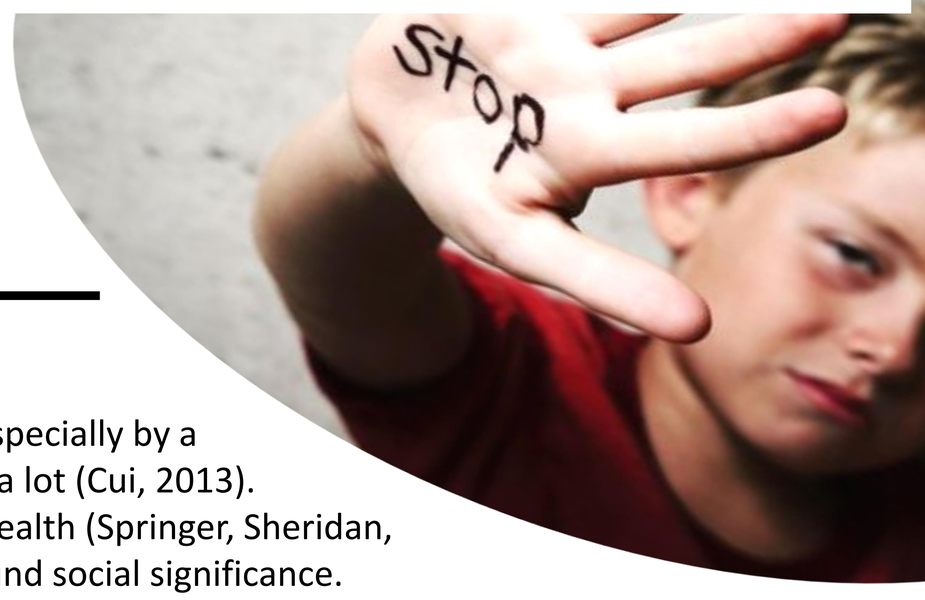
Predicting The Occurrence of Child Abuse Using Parental Data



Background

Child abuse refers to the physical, sexual, and psychological mistreatment or neglect to a child, especially by a parent or a caregiver. It is a complex social problem, from which American children also suffered a lot (Cui, 2013).

Prior literature has examined the impact of childhood abuse on long-term physical and mental health (Springer, Sheridan, Kuo, &Carnes, 2007). Therefore, the construction of prediction models for child abuse has profound social significance.

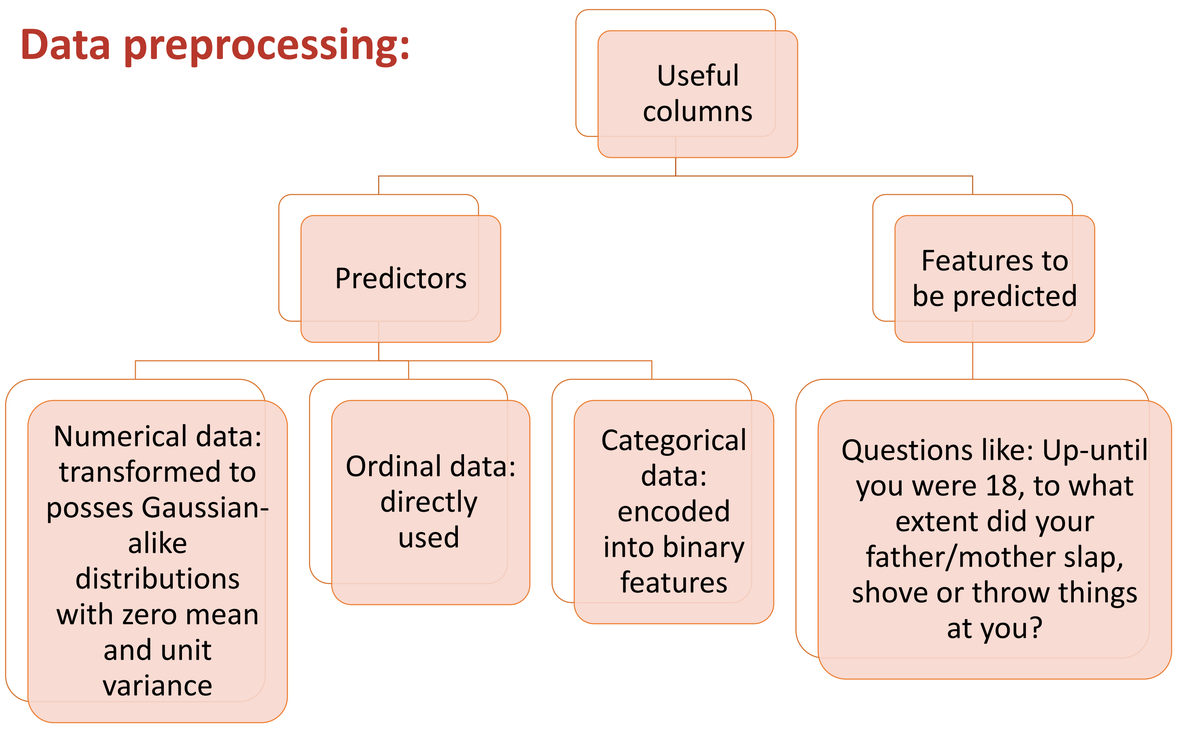
Limitations of previous research

Mainly focused on sexual abuse, physical abuse and psychological abuse are ignored	Lots of inference research, few predicting research
Mainly focused on the relationships between childhood abuse and adult performance	Sex-biased, more research on females

Methods

Dataset: Dataset used for this project is Wisconsin Longitudinal Study (WLS) dataset. The data collected from the Class of '57 respondents in 1957, 1977, 2004 were used in this project. These data provided a full record of parental information and childhood abuse experience of the Class of '57 respondents (Sewell, Hauser, Springer, & Hauser, 2003).

Data preprocessing:



Data leakage:

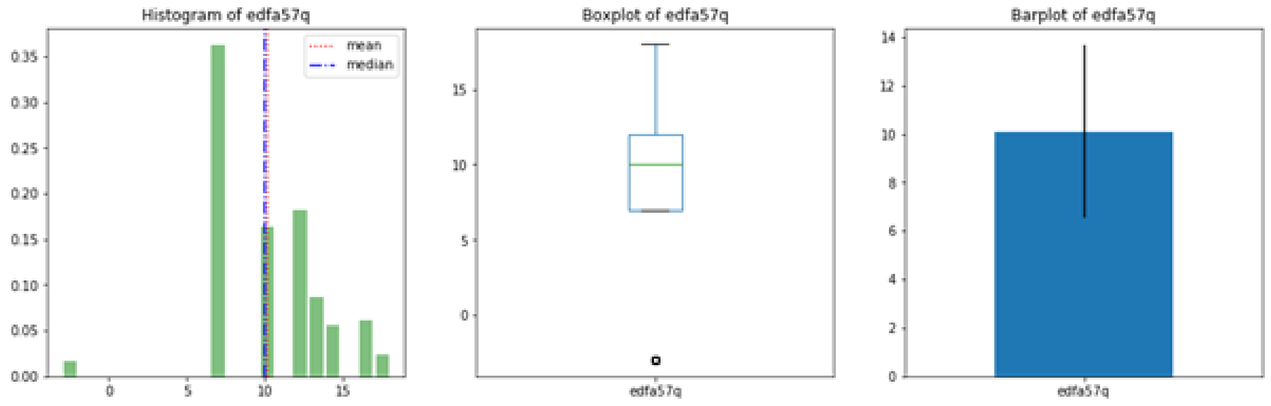
We care about the chronological order of the sample data.

- Screening our features to remove those only available after the target event happened ensures that the information indicated by all predictors is strictly before the time when respondents grew to 18 years old.

We make sure that there exists no train-test contamination.

- We did not impute missing values before generating the train-test split, so no test data was incorporated at the time of making predictions

Data statistics: The dataset size is 5931 rows × 32 columns after preprocessing. Relevant information of each variable of interest is visualized as the following (here we only use one variable as an example):



Models: First, cross-validation is used to assess the predictive performance of the models by holding out a subset of 20 percent of the data as the test set. Then, 6 models are used to predict 6 targets. Those are Decision Trees, Random Forests, Gradient Boosting, Extreme Gradient Boosting, Neural Networks, and Logistic Regression.

Results

Models Comparison: We use test accuracy as the outcome statistics to evaluate the model.

Test accuracy (%)

	decision tree	random forest	gradient boosting	XGB	neural networks	logistic regression
iv208rer	52.10	54.75	52.80	57.50	51.75	50.10
iv209rer	53.45	55.50	54.20	51.50	54.30	54.15
iv210rer	50.75	54.30	49.25	54.70	47.40	49.80
iv211rer	50.90	53.20	53.35	53.45	53.95	51.60
iv212rer	50.10	54.65	53.90	54.80	50.70	49.65
iv213rer	50.15	51.80	47.20	53.95	49.50	49.90
average	51.24	54.03	51.78	54.32	51.27	50.87

Best model analysis: We use test accuracy as the outcome statistics to evaluate the model.

Three most import features and its correlation with outcomes	Most important(correlation)	Second most important(correlation)	Third most important(correlation)
Father's verbal abuse	Age of household head(negative)	Duncan SEI score for mother(bowl shape)	NORC prestige score for father(positive)
Father's maltreatment	Mother's years of schooling(positive)	Father's years of schooling(positive)	The highest grade of school the head of household completed(positive)
Father's physical abuse	Family income rank(low)(positive)	Family income rank(lower middle)(positive)	The industry the head of household works in(technical industry) (positive)
Mother's verbal abuse	Siegel prestige score for head of household(polynomial)	Age of household head(negative)	NORC prestige score for father(positive)
Mother's maltreatment	Family income rank(low)(positive)	Mother's years of schooling(positive)	NORC prestige score for father(positive)
Mother's physical abuse	Duncan SEI score for mother(positive)	Parental income(negative)	The highest grade of school the head of household completed (inverted bowl shape)

Discussions

Operation of the best model: The operation of our best model is in line with the actual social situation--child abuse occurs not only in families with poor economic conditions and low education levels, but also in many families with high social status and good educational background.

Limitations: 1. Some potential useful predictors (e.g. the personality of the parents, whether they are addicted to smoking or alcohol) were not included in this project. 2. The predicted results are Boolean values (0 for False and 1 for True), which is relatively absolute. An optimized solution is to score the potential for abuse.

Reference list

1. 崔海英, & Cui Hai-ying. (2013). 美国虐童防控对策研究 - A Study on the Prevention and Control Countermeasures against Child Abuse in America. 政法学刊, 30(03), 99-105.
2. Sewell, W. H., Hauser, R. M., Springer, K. W., & Hauser, T. S. (2003). As we age: The Wisconsin Longitudinal Study, 1957–2001. Research in Social Stratification and Mobility, 20, 3–111.
3. Springer, Kristen W., Jennifer Sheridan, Daphne Kuo, and Molly Carnes. 2007. "Long-Term Physical and Mental Health Consequences of Childhood Physical Abuse: Results from a Large Population-Based Sample of Men and Women." Child Abuse and Neglect 31:517-530 PMCID: 17532465

Du Zhixu | 3035446851 Fan Jiarui | 3035331991 Lin Yizhou | 3035446851 Tao Yufeng | 3035447049 Zhu Zixuan | 3035449607