Predicting the occurrence of childhood abuse using parental information

COMP 3354 Statistical Learning Final Report

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

Abstract

Child abuse is a controversial social concern that merits our attention to build effective prediction models for risk control. This project leverages various statistical learning techniques to find patterns in the Wisconsin Longitudinal Study dataset and classifies whether parents are likely to mistreat their children based on their socioeconomic information. Comparing variable importance yields the conclusion that parents with lower income and higher education levels are predicted with a higher probability of abusing children.

1. Introduction

Child abuse refers to the physical, sexual, and psychological mistreatment or neglect to a child, especially by a parent or a caregiver, which is a complex social problem. In 2008, 1730 children died due to abuse in the United States. In 2012, Child Protective Services (CPS) agencies estimated that around 0.9% children in the U.S. suffered from child abuse. The shocking numbers called our attention to the child abuse problem in the U.S. (American Humane Association et al., 1976).

Most of the previous research focused on only sexual abuse among children, especially among girls (Springer, Sheridan, Kuo, &Carnes, 2007). However, general physical and psychological abuse beyond sexual mistreatment is more prevalent and yet rather ignored. (US Department of Health and Human Services, 2002, 2006). In addition, existing research established various relationships between different forms of child abuse and adult physical and mental performance (Springer, Sheridan, Kuo, &Carnes, 2003), which are typical inference research. Effective prediction models using early family information, however, do not abound. Furthermore, only less general findings were yielded in most of the previous research due to insufficient samples (Springer, Sheridan, Kuo, &Carnes, 2007).

Therefore, building predictive models based on parental information including but not limited to occupation, education and economic conditions will be the key focus of this project. We aim to better assess the potential risk of children being physically and psychologically abused by comparing various statistical learning model performances.

With a well-established predicting model, we can use the basic information of both parents to estimate whether they will have a tendency to abuse children even before their children are born. Once the risk threshold is reached, the government can provide both targeted protection to the children and proper supervision to the parents at an early stage.

2. Methods

2.1 Dataset

Dataset used for this project is Wisconsin Longitudinal Study (WLS) dataset. The WLS is a long-term study of 10,317 men and women who graduated in 1957 from Wisconsin high schools (Class of ’57 respondents) and of their randomly selected brothers and sisters (sibling respondents) (Sewell, Hauser, Springer, & Hauser, 2003). The data collected from the Class of ’57 respondents instead of siblings respondents are used in this project because of the larger size.

At the very beginning of WSL, information about parents’ education and occupation was collected in 1957. Twenty years later, in 1977, a telephone survey was carried out and more parental questions were asked. The child abuse related questions were added to the mail questinaires in 2004, when the graduates were at their 60s (Hauser, 2005). These data provided a full record of parental information and childhood abuse experience of the Class of ’57 respondents.

2.2 Sampling procedure

This section would first illustrate the prediction targets and relevant features of interest, followed by detailing how the data is sampled and preprocessed. Further treatment of the imbalanced dataset is discussed after unveiling the misleadingly satisfactory model predictive power.
Utilizing the WLS dataset, we concentrate on training models to predict possible parental physical abuse. In particular, the features to be predicted are characterized by questions like “Up-until you were 18, to what extent did your father/mother slap, shove or throw things at you?” These survey questions were answered by the graduate correspondents in retrospect and recorded in 2004 via mailing. Considering the scarcity and sensitivity of the variables pertaining to the proposed topic, we manually went through various waves of surveys in the codebooks which contain variable description and

frequency summary. Eventually, we narrow down to approximately 30 high-quality features that are ready to be fed into the model. A full view of these features and their corresponding encoding (used for visualization in later sections) could be found in the appendix. However, it is important to demonstrate the data processing mechanism in greater details, including data transformation and missing value imputation.

After the SPSS sav file obtained directly from WLS is read into a Python pandas data frame, it becomes clear that there are three kinds of data: numerical, categorical and ordinal. For instance, the parental income has a numerical meaning, while the father's nationality is categorical without any intrinsic ordering among the possible values. In contrast, the to-what-extent type of questions is ordinal and differentiates from other categorical data by indicating a clear ordering to the classes. It merits attention that our prediction targets are all ordinal. For simplicity, we treat this statistical learning problem as a classification problem. Even though we may incur the cost of losing the ordering information of each category, treating it as a regression problem on the other hand appears to bring in more inaccuracy. This is because if we were to fit regression algorithms and obtain an interval of values, no definite clues would be available as for how to divide the interval and further map to corresponding classes. Standard classifiers are therefore used in later modelling sections. In addition, deploying other tools for such ordinal regression or ranking learning problems are possible, such as the ORCA (Ordinal Regression and Classification Algorithms).

When it comes to data pre-processing, we perform standardization to numerical variables and one-hot encoding to categorical variables after imputing the missing values. Unfortunately, the dataset contains a sizable amount of missing values which are represented by survey responses such as "Don't know" or "Refused to answer". Rather than deleting any row with at least one missing value, we choose not to sacrifice the limited information but to select reasonable imputers to infer the values from what is known. Due to the experimental and controversial nature of various iterative imputers, simple imputers are used here. We then transform every numerical feature to possess Gaussian-alike distributions with zero mean and unit variance as a common practice. Categorical features are encoded into binary features and ready to be fit with models.

Furthermore, there are two more details worth noting related to data leakage, a common pitfall many practicing data scientists face. Firstly, we prevent ourselves from target leakage. It means that we care about the chronological order of the sample data. Screening our features to remove those only available

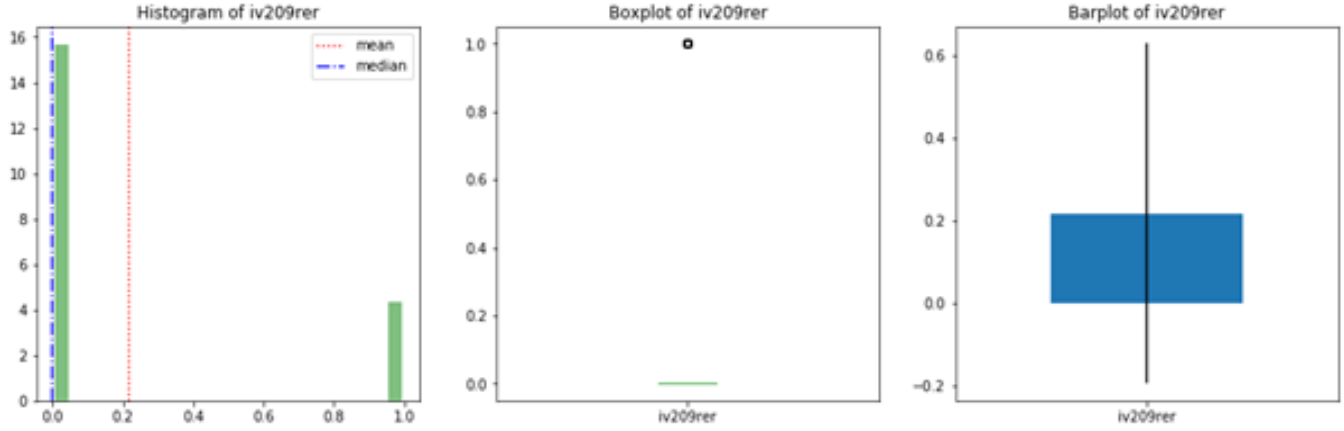
after the target event happened ensures that the information indicated by all predictors is strictly before the time when respondents grew to 18 years old. For example, how the respondents were treated in 1992-93 is well beyond their childhood and ought to be excluded. Secondly, we make sure that there exists no train-test contamination. In other words, no test data should be incorporated at the time of making predictions. This could be achieved if we do not impute missing values before generating the train-test split. Otherwise, the model may perform well even on the validation set but would fail on new data because of overfitting.

After our first trial of fitting all the subsequently introduced models, we obtain a set of seemingly promising test error rates at around 85%. Nonetheless, printing out the multi-class confusion matrices reveals that all our models predict the same category (almost no abuse) for every validation data point. This is especially common for imbalanced datasets, such as the typical machine learning problem fraudulent transaction detection. Our approach to address it is to combine over-sampling and under-sampling. Binarization is carried out to put an emphasis on distinguishing the over-represented case and the rest. In the hope of obtaining more prediction outcomes to be reflective of more severe child abuse, we resample the original dataset by adding more instances of the previously under-represented classes. The final model fitting after taking this adjustment into consideration is analyzed in later section.

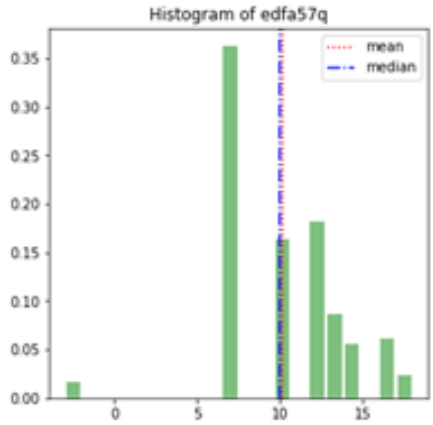
2.3 Descriptive statistics

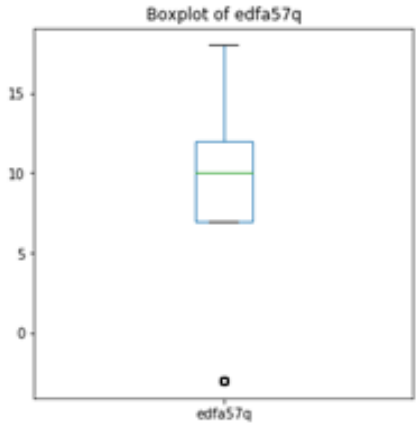
The dataset size is 5931 rows × 32 columns. Relevant information of variables of interest is summarized in the attached table in the end. We also visualize a few representative variables below.

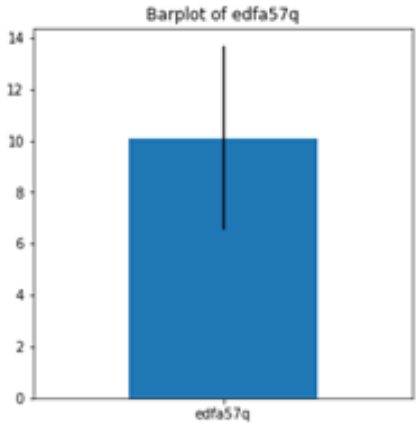
The prediction target iv209rer corresponds to the question “Up-until you were 18, to what extent did your father slap, shove or throw things at you?”, which can be easily seen as rather imbalanced.



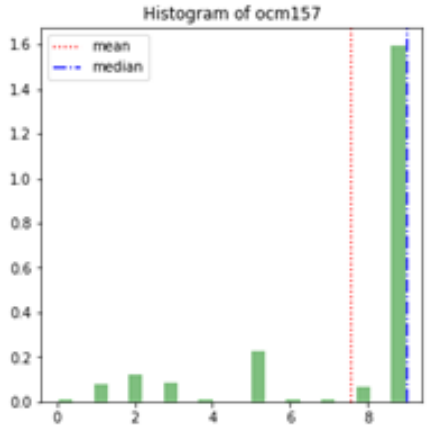
“Father's years of schooling” is an example of numerical variables, indicating that there were many who did not attend high school.

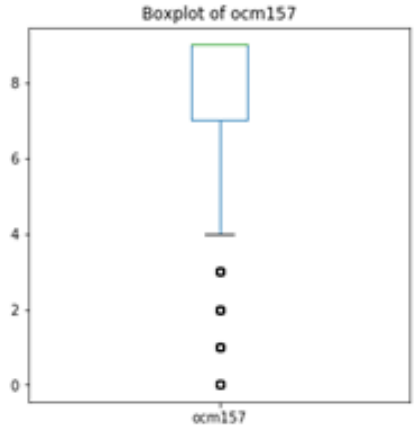


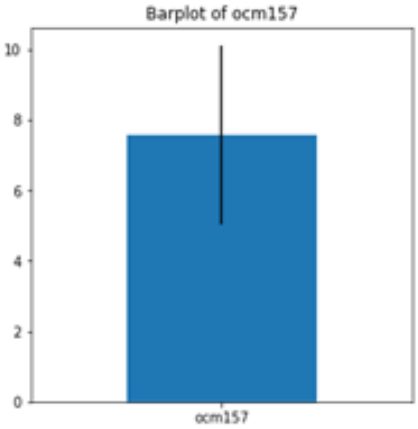




“Mother's 1957 occupation by general category.” is an example of categorical variables, which shows that most of them falls in “Student, housewife, unemployed, or occupation not reported”.







2.3 Analytical procedure

2.3.1 Statistical learning techniques

Data analyses are conducted using Python. 6 models are used to predict each of the 6 targets and the comparisons among those models are conducted. Those are Decision Trees, Random Forests, Gradient Boosting, Extreme Gradient Boosting, Neural Networks, and Logistic Regression.

Decision trees are used to predict a qualitative response, which belongs to the most commonly occurring class of training data in the region when interpreting the results of a classification tree. This method

allows us to provide a framework to quantify the values of outcomes and the probabilities of achieving them. It also gives a better visualization of the prediction results.

Random Forests construct a multitude of decision trees when training the data and output the classification. As an ensemble learning method for classification, it provides an improvement over bagged trees and the variance of which is reduced when we average the trees by combining the trees predictors such that each tree depends on the value of an independently sampled random vector. At each split of the tree, this model only considers a small subset of features rather than all the features. It gives great performance with high dimensional data and it trains the data faster than decision trees. When we average all the decision trees in random forest, the variance is averaged away so that we have a low bias.

Gradient boosting is a statistical learning technique for classification problems, which converts weak learners into strong learners and produces a prediction model to overfit a training dataset quickly. As an implementation of gradient boosted decision trees, XGB (Extreme Gradient Boosting) is a more efficient version of gradient boosting framework which contains both a linear model solver and tree learning algorithm.

We also use neural networks to generalize the linear models that perform multiple stages of processing to come to a decision, which is inspired by the structure of biological neural networks in the human brain and is designed to replicate the way humans analyze the data. This method can capture information from large amounts of data and create complex models.

Last but not the least, we analyze the data by logistic regression method. It is an appropriate regression analysis to explain the relationship between a binary or dichotomous dependent variable and one or more independent variables. This method does not require too many computational resources nor scaled input features and it is easy to interpret and regularize.

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, and 80 percent as the training set. The dataset has 5931 rows so that it has big enough test set to make a prediction on it. Without cross-validation, the training error rate might underestimate the test error rate.

2.3.2. Outcome statistics and evaluation criteria

We use test accuracy as the outcome statistics to evaluate the model. The test accuracy is all correct predictions divided by the total number of the test set, which is used for performance evaluation. The test accuracy is just calculated by (1 – test error rate). A higher test accuracy a model achieves, the better it fits this project.

Since we have 6 targets to predict in total, we may encounter a situation that different models represent different best test accuracy. To make the evaluation process more objective and more efficient, we set evaluation rules in advance. First, the best-fit model we are looking for must have the largest quantity of best prediction results among all the 6 models. Because we try to eliminate the unexpected impact of extreme results, we choose the model that gives accurate prediction in most circumstances. If more than one models fulfill this requirement, we will select the one with the largest average test accuracy, which shows the better performance overall.

3. Results

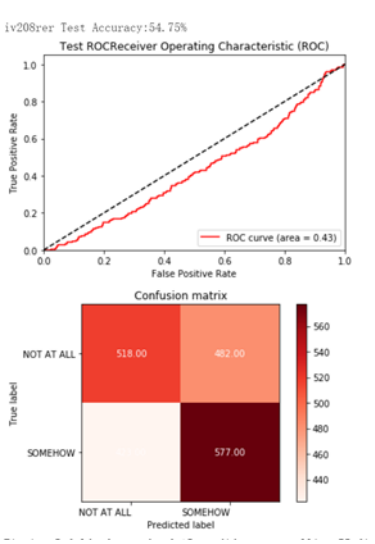
This section mainly focuses on the final performance of our prediction models, as well as the analysis of the significance of our research results.

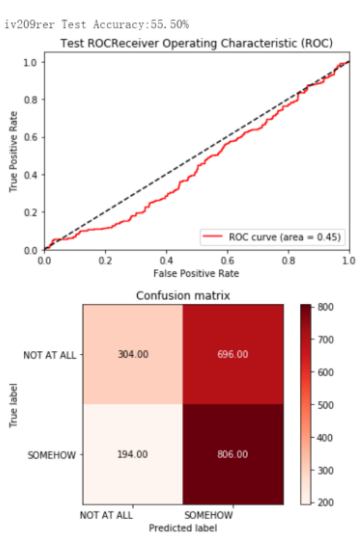
Cross Validation test error rates for our six outcomes (introduced in previous parts) predicted by different models are shown in the following table. XGB outperforms the others in terms of predicting the first, third, fifth and sixth outcomes. As for the rest, random forest and neural networks stand out with higher test accuracy. However, compared to neural networks, XGB has a more consistent performance for different outcomes, and it achieves the highest average predictive accuracy (54.32 percent). Therefore, we select XGB to predict the fourth outcome as well. Overall, our selected models are capable of obtaining a more reliable result than random guess, with all test error rates below fifty percent.

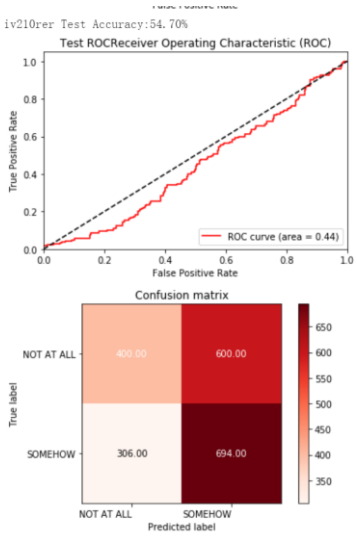
Test accuracy (%)

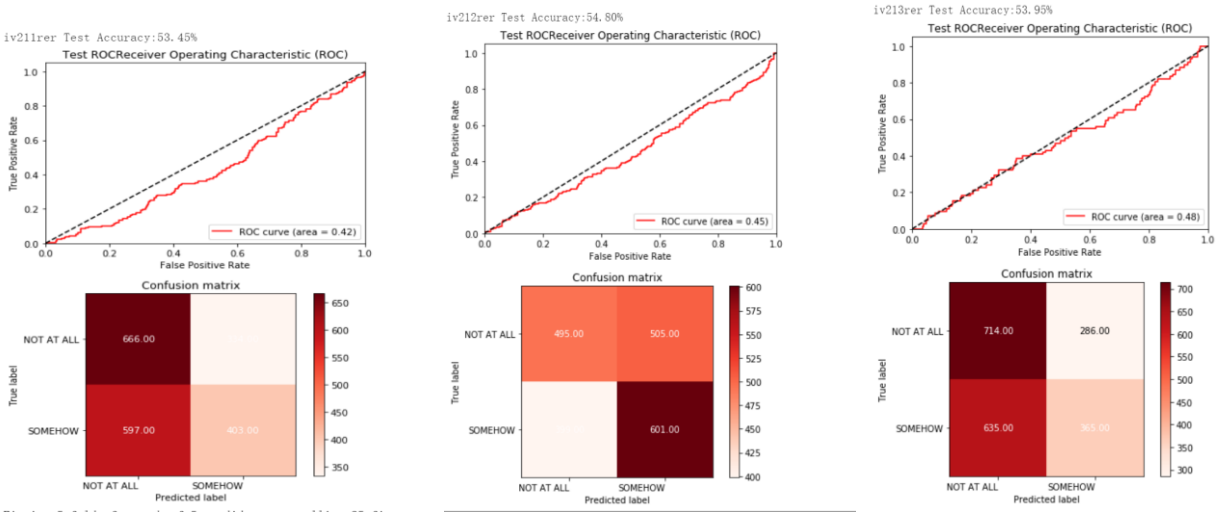
	decision tree	random forest	gradient boosting	XGB	neural networks	logistic regression
iv208rer	52.10	54.75	52.80	57.50	51.75	50.10
iv209rer	53.45	55.50	54.20	51.50	54.30	54.15
iv210rer	50.75	54.30	49.25	54.70	47.40	49.80
iv211rer	50.90	53.20	53.35	53.45	53.95	51.60
iv212rer	50.10	54.65	53.90	54.80	50.70	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

The charts below present the confusion matrices and test receiver operating characteristic plots for each of the optimal models corresponding to the respective outcomes. When predicting the second outcome (father's maltreatment), our model (random forest) tends to make "SOMEHOW" response rather than "NOT AT ALL", which leads to a higher accuracy (806 out of 1000, 80.6percent) to make predictions for positive outcomes. However, on the occasion of not being abused, random incorrectly diagnoses 304 out of 696 objects, resulting in a Negative Predictive Value (NPV) of only 30 percent. As for the other outcomes, XGB model is generally balanced to make well-judged predictions under the circumstances of either "NOT AT ALL" or "SOMEHOW", which proves the effectiveness of XGB model.



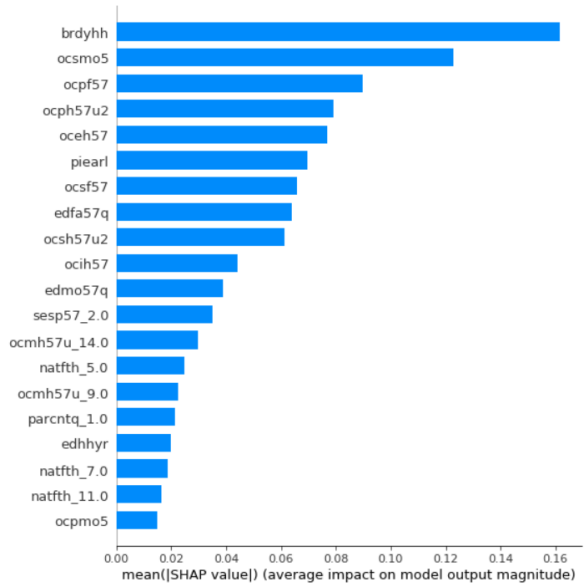


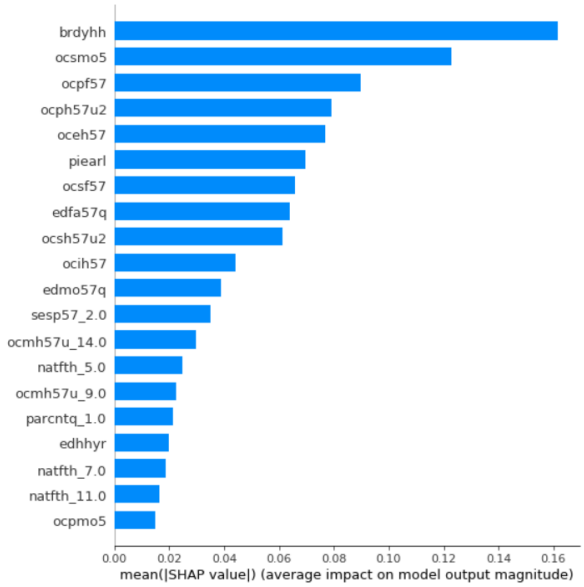


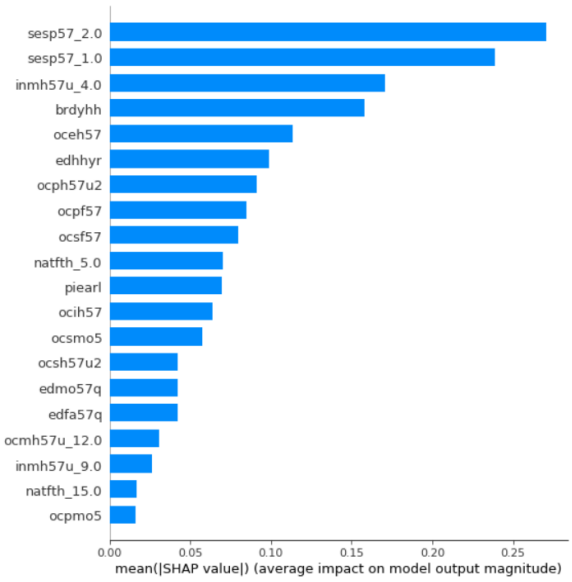


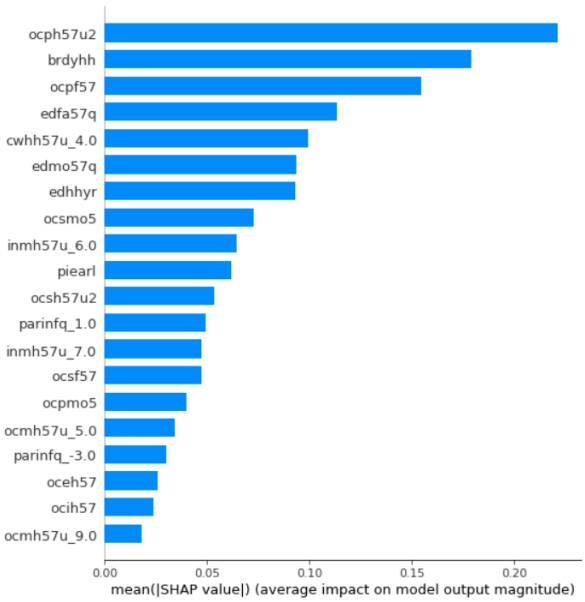
To analyze the importance of variables and interpret the models, we utilize global interpretability and partial dependency plots as follows.

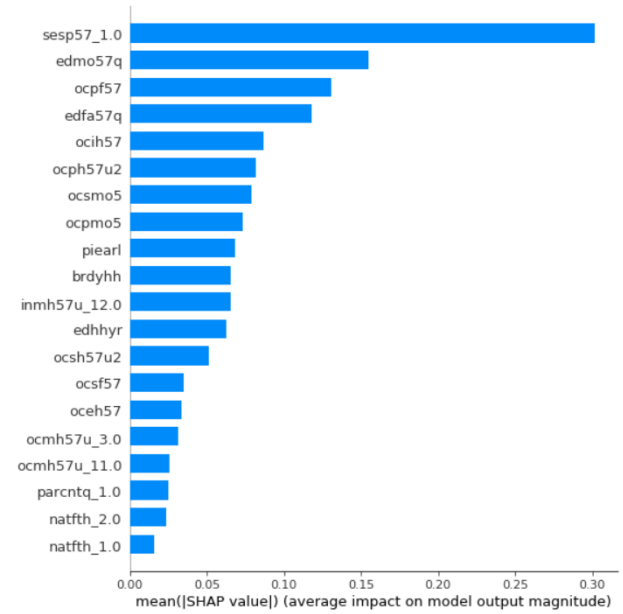
In the plots, the variable with a higher mean (|SHAP VALUE|) is more influential on the overall model performance. According to the table, we can summarize that the most significant factors are schooling, occupations, incomes and the social socioeconomic status of parents. Income is generally negatively associated with the extent to which a child has been abused. This conclusion could be explicitly demonstrated in the model for predicting the third outcome (Father's physical abuse). The positive correlations here indicate that the family with a low ranking in terms of income has a tendency to conduct physical abuse on children. However, we could also infer from the second and fifth models that parents' higher education level may result in a higher probability of childhood abuse. In addition, higher socioeconomic scores give rise to the incidence of child abuse as well. Further discussions on our findings, which may be contradicting to the common sense, will be developed in the next section.

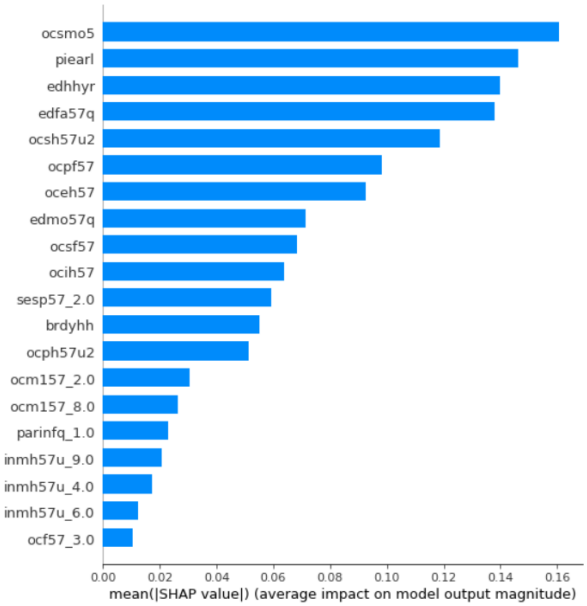

brdyhh
ocsmo5
ocpf57
ocph57u2
oceh57
piearl
ocsf57
edfa57q
ocsh57u2
ocih57
edmo57q
sesp57_2.0
ocmh57u_14.0
natfth_5.0
ocmh57u_9.0
parcntq_1.0
edhhyr
natfth_7.0
natfth_11.0
ocpmo5
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16
mean(|SHAP value|) (average impact on model output magnitude)


brdyhh
ocsmo5
ocpf57
ocph57u2
oceh57
piearl
ocsf57
edfa57q
ocsh57u2
ocih57
edmo57q
sesp57_2.0
ocmh57u_14.0
natfth_5.0
ocmh57u_9.0
parcntq_1.0
edhhyr
natfth_7.0
natfth_11.0
ocpmo5
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16
mean(|SHAP value|) (average impact on model output magnitude)


sesp57_2.0
sesp57_1.0
inmh57u_4.0
brdyhh
oceh57
edhhyr
ocph57u2
ocpf57
ocsf57
natfth_5.0
piearl
ocih57
ocsmo5
ocsh57u2
edmo57q
edfa57q
ocmh57u_12.0
inmh57u_9.0
natfth_15.0
ocpmo5
0.00 0.05 0.10 0.15 0.20 0.25
mean(|SHAP value|) (average impact on model output magnitude)


ocph57u2
brdyhh
ocpf57
edfa57q
cwhh57u_4.0
edmo57q
edhhyr
ocsmo5
inmh57u_6.0
piearl
ocsh57u2
parinfq_1.0
inmh57u_7.0
ocsf57
ocpmo5
ocmh57u_5.0
parinfq_-3.0
oceh57
ocih57
ocmh57u_9.0
0.00 0.05 0.10 0.15 0.20
mean(|SHAP value|) (average impact on model output magnitude)


sesp57_1.0
edmo57q
ocpf57
edfa57q
ocih57
ocph57u2
ocsmo5
ocpmo5
piearl
brdyhh
inmh57u_12.0
edhhyr
ocsh57u2
ocsf57
oceh57
ocmh57u_3.0
ocmh57u_11.0
parcntq_1.0
natfth_2.0
natfth_1.0
0.00 0.05 0.10 0.15 0.20 0.25 0.30
mean(|SHAP value|) (average impact on model output magnitude)


ocsmo5
piearl
edhhyr
edfa57q
ocsh57u2
ocpf57
oceh57
edmo57q
ocsf57
ocih57
sesp57_2.0
brdyhh
ocph57u2
ocm157_2.0
ocm157_8.0
parinfq_1.0
inmh57u_9.0
inmh57u_4.0
inmh57u_6.0
ocf57_3.0
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16
mean(|SHAP value|) (average impact on model output magnitude)

4.Discussion

4.1 Significance

With the advancement of cultural awareness and the improvement of civilization, people are increasingly aware that children's rights and interests should be protected, and society has a duty to meet children's basic physical and mental needs, and to provide children with as much help as possible (Lynch, 1985). However, it wasn't until Kempe and his team’s first research report on child abuse published in the 1960s that people paid serious attention and thought to this topic (Helfer &Kempe, 1968), and then realized that child abuse and neglect is a universal society phenomenon.

Based on the fact that the main abuses against children come from families, especially parents, we built models that use parental information to predict child abuse. 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.

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 has profound social significance. Whether there is a tendency to abuse children is a subjective question, which means that asking parents directly is difficult to get reliable answers. However, work, education, and economy are all objective conditions, and the results of these assessments will be more accurate and easier to obtain.

4.2 Limitations

In this section, two limitations of this project and the corresponding suggestions for further research will be discussed.

Firstly, 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. One of the most important predictors ideally is personality since competitive parents are more likely to place higher expectations on their children,

therefore are stricter (Helfer &Kempe, 1968). Nevertheless, personality is a relatively difficult variable to measure. If we try to gauge one's personality by asking himself/herself and his/her friends, it will be difficult to get objective results; if we adopt a detailed personality modeling procedure, this as a single predictor might be over-complicated. So after comprehensive consideration, we decided not to add this variable to the model for the time being. In future research, the establishment of a scientific personality evaluation system is needed, and more predictors need to be added to the model.

Secondly, the predicted results of our models are Boolean values (0 for False and 1 for True), since we directly use 'whether some particular kinds of abuse will occur' as the result. This is relatively absolute and may lead to low accurate rates. A better solution could be to estimate the probability of abuse occurring, so that the over-represented response is embodied on an interval of values rather than a single binary number.

Reference list:

1. American Humane Association, National Center on Child Abuse, Neglect, University of Denver. Center for Social Research, Denver Research Institute. Social Systems Research, Evaluation Division, ... & Abuse Reporting (US). (1976). National analysis of official child neglect and abuse reporting. American Humane.

2. 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.

3. Gil, D. G. (1970). Violence against children: Physical child abuse in the United States (Vol. 6). Cambridge, MA: Harvard University Press.

4. Hauser, R. (2005). Survey Response in the Long Run: The Wisconsin Longitudinal Study. Field Methods, 17(1), 3-29.

5. Helfer, R. E., & Kempe, C. H. (Eds.). (1968). The battered child (p. 110). Chicago: University of Chicago Press.

6. Johnson, W., & l'Esperance, J. (1984, July). Predicting the recurrence of child abuse. In Social Work Research and Abstracts (Vol. 20, No. 2, pp. 21-26). Oxford University Press.

7. Lundahl, B. W., Nimer, J., & Parsons, B. (2006). Preventing child abuse: A meta-analysis of parent training programs. Research on Social Work Practice, 16(3), 251-262.

8. Lynch, M. A. (1985). Child abuse before Kempe: An historical literature review. Child abuse & neglect, 9(1), 7-15.

9. orca: Ordinal Regression and Classification Algorithms, AYRNA, 2017-11-21, retrieved 2017-11-21

10. 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.

11. Springer, K. W., Sheridan, J., Kuo, D., & Carnes, M. (2003). The long-term health outcomes of childhood abuse: An overview and a call to action. Journal of General Internal Medicine, 18, 864–870.

12. Springer, K. W., Sheridan, J., Kuo, D., & Carnes, M. (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

13.US Department of Health and Human Services. (2002). Child maltreatment 2000. Washington, DC: US Government Printing Office. Retrieved September 2006 from http://www.acf.hhs.gov/programs/cb/pubs/cm00/index.htm.

14. US Department of Health and Human Services. (2006). Child maltreatment 2004. Washington, DC: US Government Printing Office. Retrieved September 2006 from http://www.acf.hhs.gov/programs/cb/pubs/cm04/index.htm.

Table:

variable code	survey question	categ/ num/ord	mean	std	missing values
Response Variable					
iv208rer	*Up-until you were 18, to what extent did your father insult or swear at you?*	ordinal	1.3594	0.7101	-3:REFUSED 224 -1:DON'T KNOW 1
iv209rer	*Up-until you were 18, to what extent did your father slap, shove or throw things at you?*	ordinal	1.2935	0.6238	-3:REFUSED 242 -1:DON'T KNOW 2
iv210rer	*Up-until you were 18, to what extent did your father treat you in a way that you would now consider physical abuse?*	ordinal	1.1948	0.5952	-3:REFUSED 240 -1:DON'T KNOW 1
iv211rer	*Up-until you were 18, to what extent did your mother insult or swear at you?*	ordinal	1.2059	0.5605	-3:REFUSED 347
iv212rer	*Up-until you were 18, to what extent did your mother slap, shove or throw things at you?*	ordinal	1.2416	0.5571	-3:REFUSED 349
iv213rer	*Up-until you were 18, to what extent did your mother treat you in a way that you would now consider physical abuse?*	ordinal	1.1149	0.4594	-3:REFUSED 345
Predictor Variable					
parcntq	*To what extent have you discussed your plans with your parents?*	ordinal	2.5646	0.7511	-3:REFUSED 122
parinfq	*How much did your parents influence your plans?*	ordinal	2.1209	0.9429	
edfa57q	*Father's years of schooling*	numerical	10.1111	3.5708	-3:REFUSED 247 -1:DON'T KNOW 582
edmo57q	*Mother's years of schooling*	numerical	10.4398	3.4598	-3:REFUSED 235 -1:DON'T KNOW 476
sesp57	*How does your family income or wealth compare with families in your community?*	ordinal	2.9997	1.1250	-3:REFUSED 313
piearl	*Earliest available parental income in hundreds of dollars.*	numerical	53.9867	55.8542	-3 :No information 1346
ocf57	*Father's occupation by general category*	categorical	2.4815	1.4534	
ocsf57	*1950 Duncan SEI score for father's 1957 occupation.*	numerical	26.8829	23.5814	-3 :No information 1346
ocpf57	*1950 NORC prestige score for father's 1957 occupation.*	numerical	53.3393	23.0332	-3 :No information 1346
ocm157	*Mother's 1957 occupation by general category.*	categorical	7.5754	2.5451	
ocsmo5	*1950 Duncan SEI score for mother's 1957 occupation.*	numerical	9.4017	18.8777	
ocpmo5	*1950 NORC prestige score for mother's 1957 occupation*	numerical	17.4583	28.6535	
bklvpr	*Did you live with both parents most of time up until 1957?*	categorical	1.0697	0.2862	-3 :NOT ASCERTAINED IN BOTH 1975 AND 1992/93 INTERVIEWS 718
bkhs57	*Summary code for head of household in 1957.*	categorical	1.0801	0.3831	-3 :NOT ASCERTAINED IN BOTH 1975 AND 1992/93 INTERVIEWS 718
brdyhh	*Age of 1957 head of household in 1975.*	numerical	64.1777	16.2016	-3 :NOT ASCERTAINED 1206 -1 :DON'T KNOW 134
edhhyr	*Highest grade of school the 1957 head of household completed.*	numerical	9.7934	3.4956	-3 :NOT ASCERTAINED -1 :DON'T KNOW 467
bkhhw5	*Did the head of household usually work in 1957?*	categorical	0.8650	0.8655	-3 :NOT ASCERTAINED 1200 -1 :DON'T KNOW 8
cwhh57u	*Class of worker code for head of household in 1957.*	categorical	2.1090	1.3894	-3 :NOT ASCERTAINED -2 : Inappropriate 157 -1 : DON'T KNOW 45
inmh57u	*1970 Major industry code for head of household in 1957.*	categorical	4.5520	2.8874	-3 :NOT ASCERTAINED -2 : Inappropriate 157 -1 : DON'T KNOW 43
ocmh57u	*1970 Major occupation code for head of household in 1957.*	categorical	9.4183	4.8134	-3 :NOT ASCERTAINED -2 : Inappropriate 157 -1 : DON'T KNOW 43
ocsh57u2	*1970 Duncan SEI score for head of household in 1957*	numerical	346.9955	234.2402	-3 :NOT ASCERTAINED -2 : Inappropriate 157 -1 : DON'T KNOW 53
ocph57u2	*1970 Siegel prestige score for head of household in 1957*	numerical	398.8961	121.2183	-3 :NOT ASCERTAINED -2 : Inappropriate 157 -1 : DON'T KNOW 53
oceh57	*1970 Occupational education score for head of household in 1957.*	numerical	211.3751	222.9896	-3 :NOT ASCERTAINED -2 : Inappropriate 157 -1 : DON'T KNOW 53
ocih57	*1970 Occupational income score for head of household in 1957.*	numerical	294.2264	203.4347	-3 :NOT ASCERTAINED -2 : Inappropriate 157 -1 : DON'T KNOW 53
natfth	*Father's nationality.*	categorical	11.4545	6.2368	-3 :NOT ASCERTAINED 744 -1 :DON'T KNOW 120
zparnf	*When you were a high school senior did your parents want you to go to college?*	ordinal	1.3406	1.1473	-3 :NOT ASCERTAINED 1236 -1 :DON'T KNOW 10