MSSP 608: Practical Machine Learning Project Report – Child Abuse Risk Prediction

Google Colab Notebook:
https://colab.research.google.com/drive/1kDC5rokfpwjf_43IW0PuX2CW_xGVKNBl?usp=sharing

Part 1: Project Narrative

Background

In an era in which child protective service (CPS) agencies face increased demands on their time and in an environment of stable or shrinking resources, great interest exists in improving risk assessment and decision support. There are 6-7 million children who are reported for alleged abuse or neglect every single year in the US (U.S. Department of Health and Human Services, 2016). The consequent of not making the right decision for those children in danger can be fatal. Yet, the resources of child welfare agency are limited, they cannot send social workers to all the family that is reported. Thus, accurately recognize the risk and efficiently allocate resources is important.

Predictive risk modeling (PRM) has been adopted to assess children's risk using administrative data. Most recently, Allegheny County, Pennsylvania, has implemented a PRM tool in its CPS hotline known as the Allegheny Family Screening Tool. The tool uses data from 27 departments housed in the county's data warehouse. The PRM model produces a risk score, enabling hotline workers to determine if referrals should be screened in for investigation (Vaithianathan, Jiang, Maloney, & Putnam-Hornstein, 2016a, 2016b). Risk scores (ranging from 1 to 20) are assigned to the entire household, not just the child victim, indicating the likelihood of a placement or rereferral in the 365 days following the hotline call. Further work is currently underway to explore models that might be deployed further upstream, helping to prioritize families for various early intervention and family support programs.

Stakeholders

- Victims. The misclassification of risk level has direct impact on maltreated children. Overrating the risk level (e.g. removing a child from a safe family) would cause extra harm to the child since the family was broken arbitrarily. Underrating the risk level (e.g. leaving the victim to a dangerous environment), on the other hand, can be fatal.
- Child protective services (CPS). The automation can improve the efficiency of child protective agencies. Misclassification can cause the misallocation of resources and thus lower the efficiency of agencies. Extreme misconduct can invoke legitimate problem.

Dataset Description

Data source. The dataset was retrieved from National Child Abuse and Neglect Data System (NCANDS). The dataset contains case-level information of children that are already identified as victims from 1998-1999. NCANDS collected data annually from those child protective services agencies able to provide electronic child abuse and neglect records. The newest dataset was published in 2018, data of year 1999 is the most recent data that can be acquired for coursework use.

Data size. The raw dataset contains 227,064 cases and 62 variables. Variables includes basic information concerning the report and the child, information about the type of maltreatment, the support services provided to the family, and any special problems that were identified for the child, caretaker, or family. After filtering prospective variables and leave out missing data, there left 30,415 cases and 21 variables in the dataset.

Predicted variables. There are in total 21 variables containing several aspects of reported cases. Most of the variables are in binary format. When an alleged child abuse is reported (case information, see Table 1), child protective agency will send social workers to investigate the case. During the process of investigation, information of the child and the family is collected (child/family problem, see Table 1). Eventually the agency will find out whether and what maltreatment the child received (case result, see Table 1).

Table 1. Summary of Predicted Variables

Subset	Variable	Format
Case Information	State	String
	Report Date	Date
	Disposition Date	Date
	Report Source	Number (1-11)
Case Result	Child Victim of Physical Abuse	Binary (0, 1)
	Child Victim of Neglect	Binary (0, 1)
	Child Victim of Sexual Abuse	Binary (0, 1)
	Child Victim of Psychological Abuse	Binary (0, 1)
Child Demographic Information	Age	Number (0-21)
	Sex	Binary (1, 2)
	Race	Number (1-4)
	Prior Victimization Status	Binary (1, 2)
Child Problem	Child Problem, Drugs or Alcohol	Binary (1, 2)
	Child Problem, Mental or Physical	Binary (1, 2)
Family Problem	Caretaker Problem, Drugs or Alcohol	Binary (1, 2)
	Caretaker Mental or Physical Problem	Binary (1, 2)
	Inadequate Housing	Binary (1, 2)
	Financial Problem	Binary (1, 2)
	Public Assistance	Binary (1, 2)

Part 2: Primary Task

Task Description

The primary task aims to train a model to predict future risks of maltreated children, using administrative data from NCANDS.

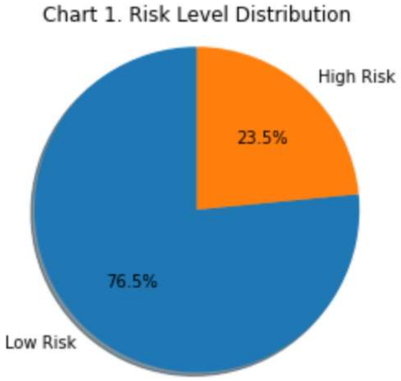
Outcome variable. When CPS believes that an incident of child abuse or neglect has happened, several outcomes can occur:

- **Low risk**/No further action. The reported child maltreatment was considered a one-time incident, the child is considered to be safe, and there is no or low risk of future maltreatment.
- **High risk**/Social Support or Child removal. There is a risk of future maltreatment, the family may be offered in-home services to reduce that risk and strengthen the family's protective capacities. If the child has been seriously harmed, the child is considered to be at high risk of serious harm, or the child's safety is threatened, the agency will remove the child.

Risk level is the outcome variable, which was generated from two binary variables: family support and foster care. Outcome labels and definition can be found in Table 2. About 77% of the cases were considered low level, which means no further actions need to be taken. Meanwhile, high risk cases account for 23% and need immediate actions.

Table 2. Summary of Outcome Variable

Variable	Label	Format	Standard	Count
Risk Level	Low Risk	String	Family Support Services not Provided, Foster Care Services not Provided	23274 (77%)
	High Risk	String	One or both Family Support Services and Foster Care Services Provided	7141 (23%)



Experimental Setup

Before building up the model, the whole dataset was divided the into three subsets: training set (80%), development set (10%), and test set (10%). Training set was used for training in every step of optimization; the test set was used for testing in optimizations and was merged to training set in the final model. The 10% test set was not used until the testing in final model.

Metrics

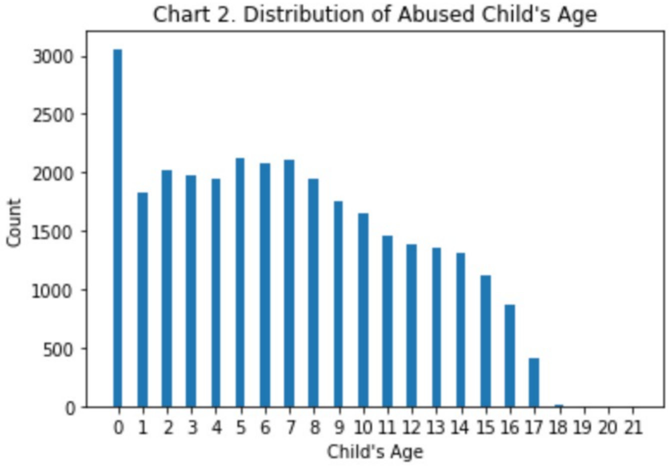
Risk prediction on child abuse is a high-stakes human decision-making process. Although the risk prediction is a reference to human decision, it is human that makes final call, the model should reach at least 80% accuracy. Recall in this model is also important and should reach at least 0.7, since assigning low risk to a high-risk case is riskier than assigning high risk to a low-risk case.

Features

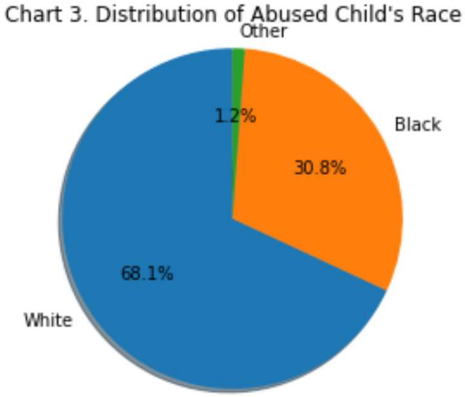
Descriptive Statistics

The following charts provide descriptive analysis of the features included in the initial model.

- **Age.** In this dataset, the age of abused child ranges from 0 to 21. Children below 1 year old are the most victimized cohort, children below 7 years old are also generally more maltreated. Pre-school age takes up higher proportion of child abuse cases. As age grows, the case of child abuse declines.



- **Race.** White and black children take up 68% and 30% of the whole dataset, respectively. Other races include Native American and Asian American, etc. In this project, the race feature includes white and black and leaves out the rest.



- **Type of Maltreatment.** Maltreatment can be divided into four categories: physical abuse, neglect, sexual abuse and phycological abuse. Dividing by category, neglect (failing to provide for a child's basic needs) takes up the highest proportion, physical abuse comes to the second, then comes sexual abuse and phycological abuse.

It is also common that multiple maltreatments were witnessed in one case. Thus, in the process of feature optimization, maltreatment features were also transformed into one case feature, indicating the how many maltreatments were seen in one case. Chart 4-2 shows that single abuse is the most common circumstance. Double maltreatments take up small proportion of the cases, while triple and quadruple maltreatments are rare.

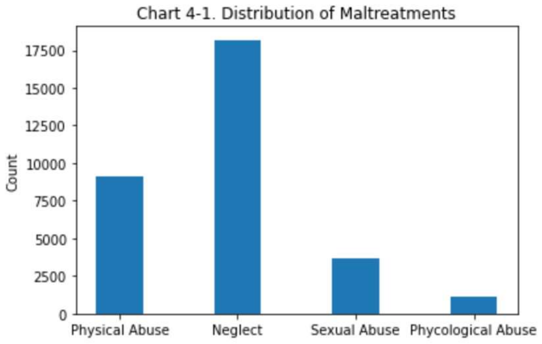


Chart 4-1. Distribution of Maltreatments

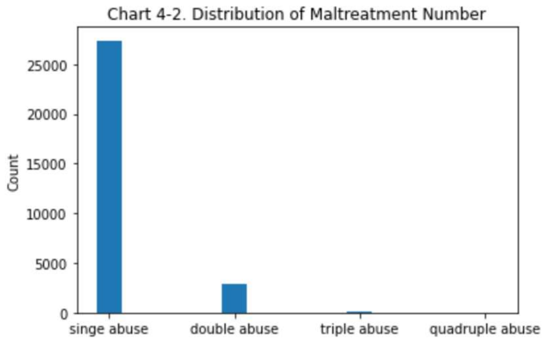


Chart 4-2. Distribution of Maltreatment Number

- **Report Source.** Abuse cases were reported from various sources. About 20% cases were reported from legal personnel. Educational personnel, friends/neighbor, medical personnel and social services were also major sources.

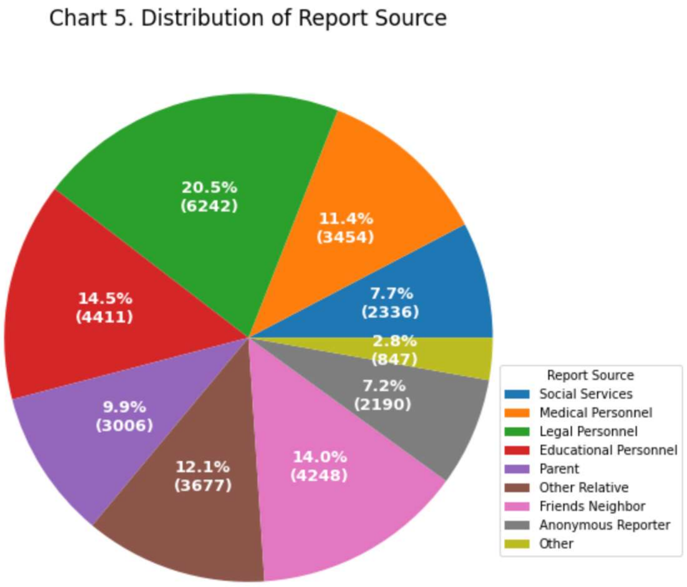


Chart 5. Distribution of Report Source

- **Report Month.** The month that cases were reported generally spread evenly across the whole year, with August the highest and December the lowest. In the model training part, **Process Time.** In the model optimization part, process_time was generated from report date and disposition date, representing how long did it take to process a case. Nearly half of the cases were processed within one month, most of the cases were processed within 3 months. There were still cases that took a long time to process, even longer that a year.

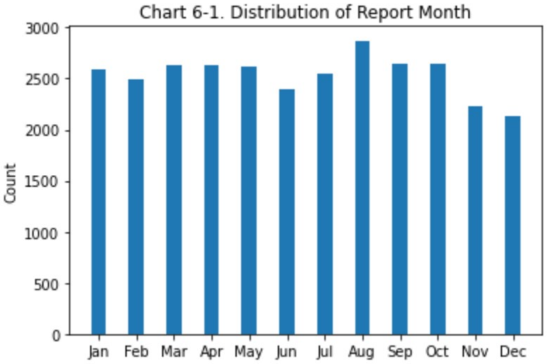


Chart 6-1. Distribution of Report Month

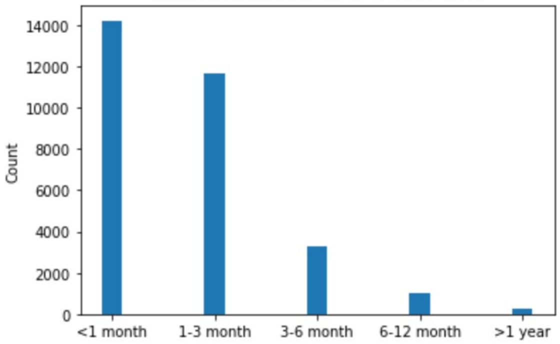


Chart 6-2. Distribution of Process Time

Feature Optimization

With all selected features included in the initial model without tuning, several adjustments were made in the process of feature optimization. **Table 3.** provides how features changed in each step and how performance of the model improved.

- Step1: Use report_date and disposition_date to generate a new variable: process_time, it represents how many days did it take to process a case.
- Step2: Combine four case result features into one. The new feature case_result has four labels: single abuse, double abuse, triple abuse and quad abuse.
- Step3: Transform age to age group: <1, 1-7, 8-16, >17.
- Step4: Remove three features: age, sex, and prior victimized status.
- Final feature set: state, report month, process time, report source, race, case result, child's drug/alcohol issue, child's mental/physical issue, caregiver's drug/alcohol issue, caregiver's mental/physical issue, house problem, financial problem, public assistance.

Table 3. Summary of Feature Optimization Process

	Operation	Feature	Accuracy	Kappa	Recall
Initial Model		all features	71.8	0.241	0.435
Step 1	add	process_time	77.2	0.383	0.539
Step 2	combine	case_result	76.6	0.369	0.533
Step 3	transform	age	79.2	0.436	0.580
Step 4	delete	sex/age/prior	82.5	0.519	0.627

Classifier

Decision Tree Classifier was adopted in the initial model. The initial setting of hyperparameters is by default. The model was trained on training set and tested on development set. **Table 4** provides how hyperparameters changed and the change of performance.

- Step 1: Compare two classifiers: Decision Tree and Logistic Regression Model.
- Step 2: Tune hyperparameters of the better performance classifier (Decision Tree).
    - ➢ Criterion: entropy (default), gini
    - ➢ Min impurity decrease: 0 (default), 0.0001, 0.001
    - ➢ Min samples split: 2 (default) to 20

- Final Model: Decision tree classifier with default hyperparameter settings.

Table 4. Summary of Hyperparameter Tuning

	Classifier	Hyperparameters			Accuracy	Kappa	Recall
		criterion	min impurity decrease	min samples split			
Best Feature	DecisionTree	entropy	0	2	82.5	0.519	0.627
Step 1	Logistic	*	*	*	76.4	0.145	0.627
Step 2	DecisionTree	entropy	0	2	82.5	0.519	0.627

Result

The final model was trained from (training + development) set and tested on the 10% test set that was never used before. Other than the confusion matrix, **Table 5** provides performance of the final model. Overall, the model reached the goal of 80% accuracy. The 0.7 recall is also satisfying. On the other hand, Kappa is below 0.7 and is not good enough. Since accuracy and recall are prioritized in this model, the result is overall qualified.

- Classifier: Decision Tree
- Training Set: 80% training set + 10% development set
- Test Set: 10% test set
- Feature Set: state, report month, process time, report source, race, case result, child’s drug/alcohol issue, child’s mental/physical issue, caregiver’s drug/alcohol issue, caregiver’s mental/physical issue, house problem, financial problem, public assistance.
- Hyperparameters: default
    - Criterion: entropy
    - Min impurity decrease: 0
    - Min samples split: 2

Table 5. Performance of the Final Model

Accuracy	Kappa	Recall
84.1	0.65	0.699

Confusion Matrix

		Predicted	
		low	high
Actual	low	500	215
	high	269	2058

Error Analysis

With 84.1% accuracy, there are in total 484 cases in the test set that the model assigned the wrong risk level. Certain pattern can be seen in the error cases. The distribution of the following features and labels are difference from those distribution in the test set.

- **Child mental/physical problem.** The current model is likely to make mistakes on abused children with mental or physical problem. In the test set, 10% of the cases are children with mental or physical problem, while in the error set there are 18%.

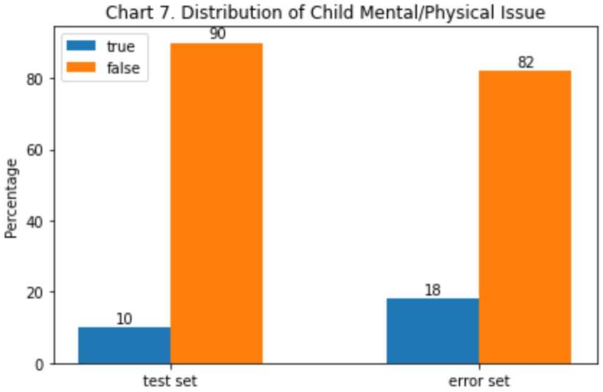


Chart 7. Distribution of Child Mental/Physical Issue

- **Case result, single/double.** The feature case result showed different distributions in test set and error set. Specifically, cases regarding single abuse and double abuse are the circumstances that current model made more mistakes.

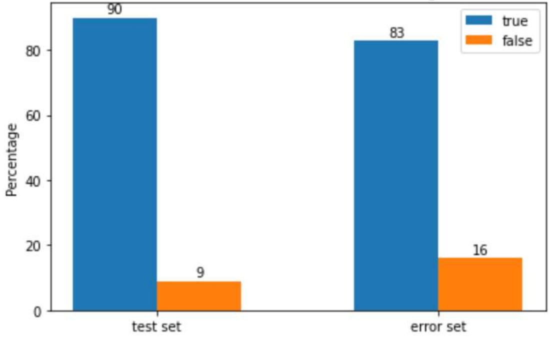


Chart 8-1. Distribution of Case Result - Single Abuse

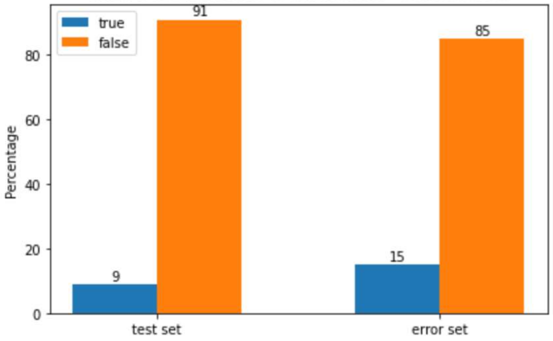


Chart 8-2. Distribution of Case Result - Double Abuse

- **Report month, July.** The error cases showed higher proportion in the report month of July than in the test set, which indicates that the model made slightly more mistakes to cases reported in July than other months of the year.

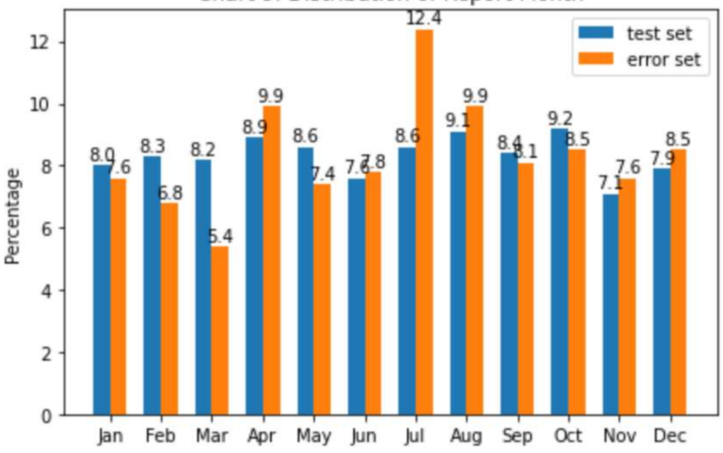


Chart 9. Distribution of Report Month

Part 3: Extension Task

Task Definition

The application of risk prediction model in child welfare suffers from fairness concerns. The extension task aims to conduct a fairness audit of the dataset, specifically on demographic features: age and race.

Motivation

According to Children's Data Network in University of Southern California, in their system, poor and minority communities are over-represented in data collected by counties. Thus, the data is already biased against certain group of communities and those pre-existing biases might be a legitimate concern.

The disparity between races, age groups and sex are irresponsible for abused children of all subgroups. Unlike other predictive algorithm such as crime prediction, which may assign higher risks to African American people and let go criminals from other races, overrating risks of African American children means underrating risks for children of other races that can lead to worse consequences to those children.

Methods

The fairness of this dataset will be examined on demographic features **race** (black and white) and **age**. Feature age was transformed to a binary feature, pre-school (below 7 years old) and non-preschool (above 7 years old), based on children's average age in the test set. The fairness is measured by the following methods:

- **Demographic Parity**. Calculate distribution of risk levels of each subgroup. The model is considered fair if similar proportions of each subgroup get assigned to different risk levels.
- **Prevalence**. Calculate baseline distributions of risk levels assigned by the model and those distributions divided by age and race.
- **Error Rate Parity**. Calculate false positive rate (FPR) and false negative rate (FNR) of all subgroups. Fairness would be indicated by similar FPR and FNR of each subgroup.

Results

- **Demographic Parity.** From the charts below, we can see that all race groups and age groups get the same distribution of risk labels from the classifier. Specifically, white and black children in the dataset have similar proportions of low and high risk levels. Preschool children and school children also have similar proportions of both risk levels.

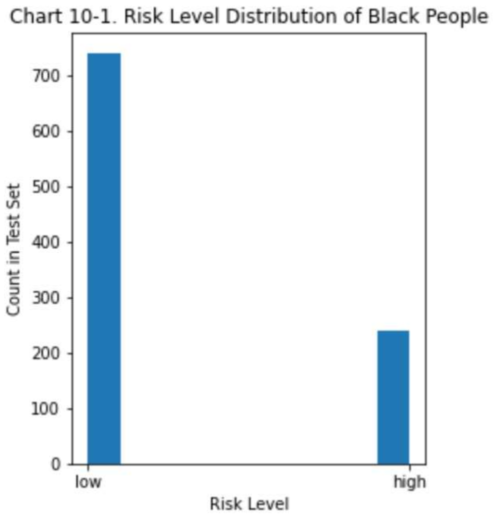


Chart 10-1. Risk Level Distribution of Black People

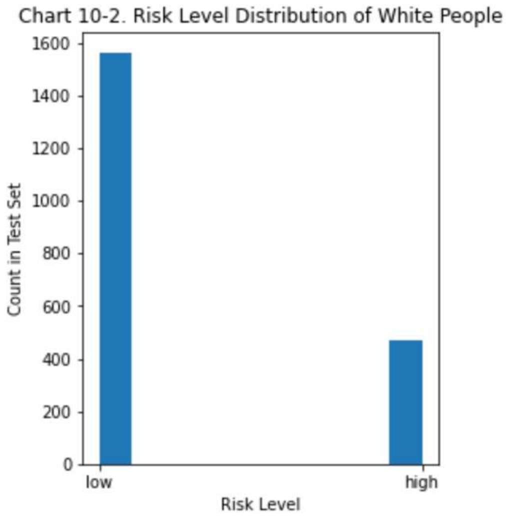


Chart 10-2. Risk Level Distribution of White People

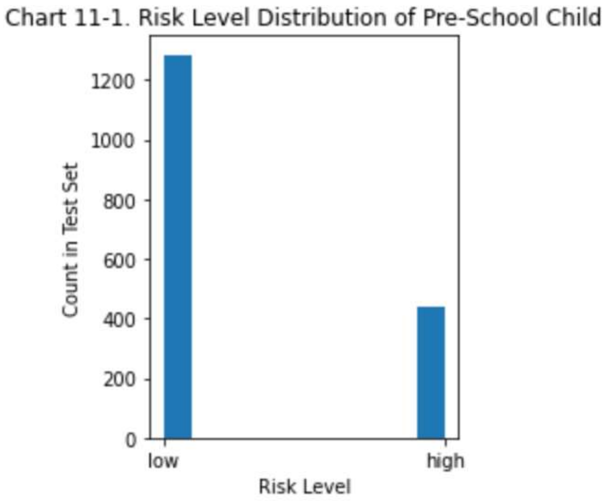


Chart 11-1. Risk Level Distribution of Pre-School Child

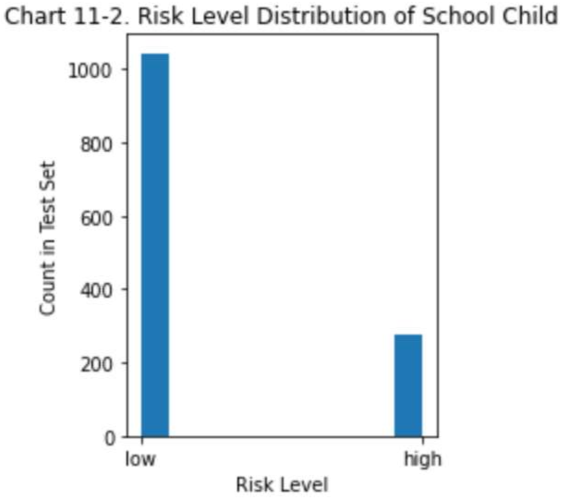


Chart 11-2. Risk Level Distribution of School Child

- **Prevalence.** Similar to demographic parity, the risk level distribution is considered even between age groups and race groups in the dataset, as shown in the charts below.

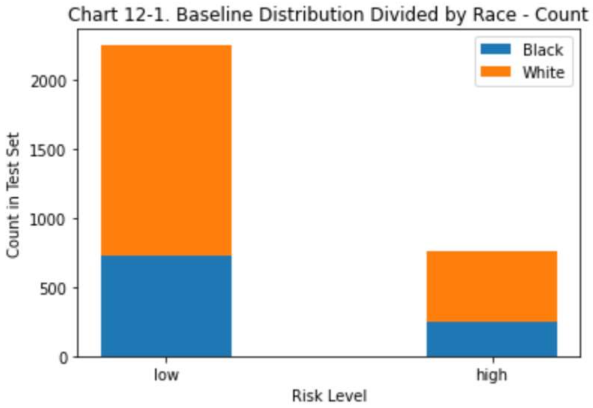


Chart 12-1. Baseline Distribution Divided by Race - Count

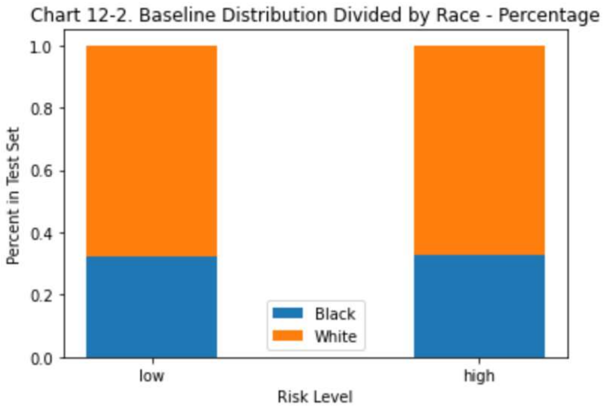
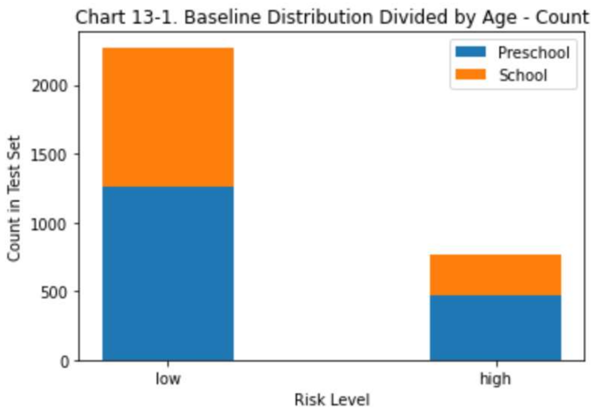
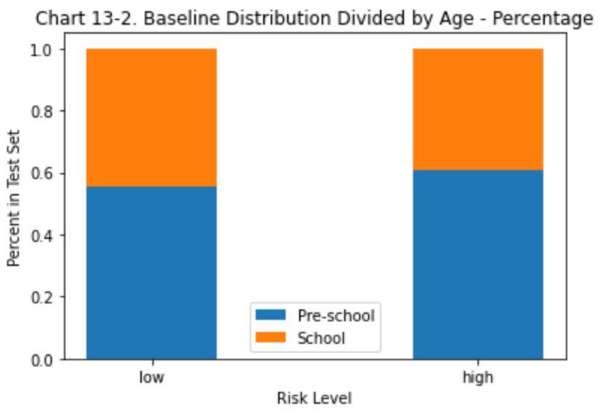
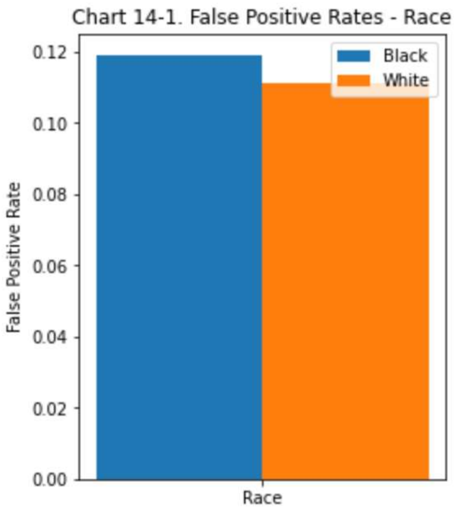


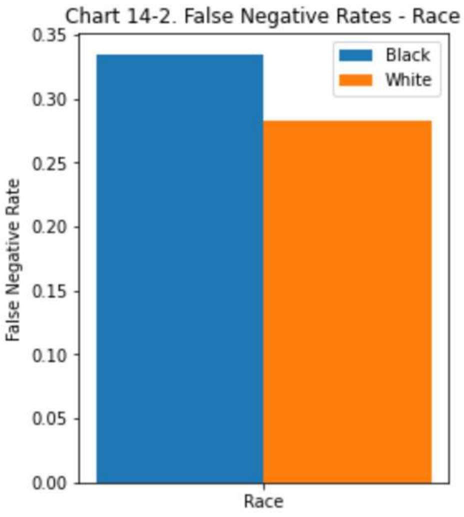
Chart 12-2. Baseline Distribution Divided by Race - Percentage

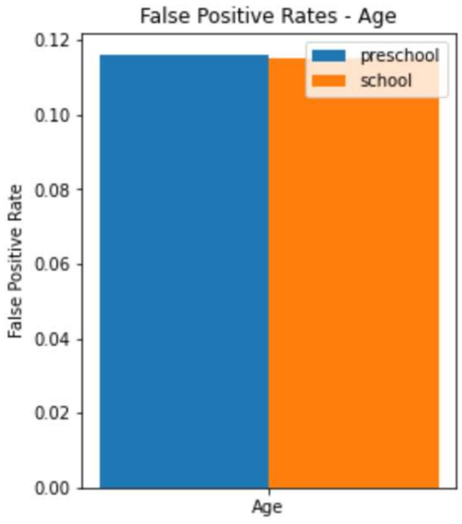




- **Error Rate Parity.** Looking at race groups, both FPR and FNR are higher for black children, though not substantiated (less than 0.5% difference). In the age group, pre-school and school children share same FPR while FNR is slightly higher for school children. **Table 6** provides error rates in detail.







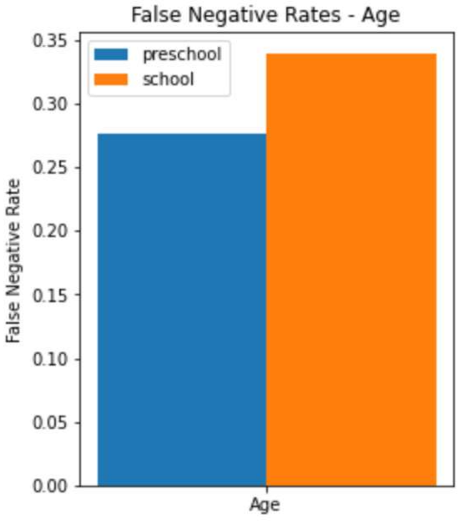


Table 6. Error Rates by Group

	Subgroup	FPR*	FNR**
Race	Black	0.12	0.33
	White	0.11	0.28
Age	Pre-school	0.12	0.28
	School	0.12	0.34

* False Positive: a low-risk case was assigned high risk by the model

** False Negative: a high-risk case was assigned low risk by the model

- **Conclusion.** With the quantitative evidence above, the dataset is considered fair overall in terms of race and age. However, the fairness of the child abuse dataset only represents data in the 2000's and does not guarantee the same result of the nowadays data.

Project Summary

Conclusion

The project successfully trained a model that predict risk levels of maltreated children based on NCANDS data of year 1999. After optimizing features and hyperparameters, the model reached 84.1% accuracy, which is a qualified result. As for the fairness, this specific dataset does not discriminate against any race or age group.

Limitation

- **Outdated dataset.** Due to the restrictions of NCANDS, year 1999 is the most recent data that can be acquired for coursework. With the nature of the data changing over time, the model should be adjusted and updated for future use.
- **Geography limitation.** In year 1999, 10 out of 50 states were recorded in the dataset. After leaving out cases with missing data, only 3 states were included in the training set. The most recent dataset contains all 50 states and could show different results in this model.
- **Performance.** The accuracy (84.1%) and recall (0.7) of the model are important and still has space to improve. As a referencing tool for child protective agencies, the model would be more reliable with 95% accuracy and 0.9 recall.

Reference

U.S. Department of Health and Human Services. (2016). Child maltreatment 2014. Retrieved from https://www.acf.hhs.gov/sites/default/files/cb/cm2014.pdf.

Vaithianathan, R., Jiang, N., Maloney, T., & Putnam-Hornstein, E. (2016a). Implementation of predictive risk model at the call centre at Allegheny County.

Vaithianathan, R., Jiang, N., Maloney, T., & Putnam-Hornstein, E. (2016b). Developing predictive risk models at call screening for Allegheny County: Implications for racial disparities.