Predicting Foster Child Exit Outcomes

Jason Huang
Stanford University, Department of Economics

Introduction

Currently, there are around 67,000 children in the California foster care system. Foster care system is a government run system in which a minor is placed into an institution, group home, or a private home. The reason a child enters the foster care system may vary, including abuse and neglect by or the death of the parents. The median stay in California

The child can exit from foster care in different ways, and this project focuses on 4 main ones: reunification with the primary family, adoption by another family, aging out of foster care, and running away. This project aims to predict the exit outcome based on the child's health state, reasons for being removed from the child's immediate family, and the length of stay in foster care.

This project is only a predictive exercise, and I cannot claim any causalities. For instance, I found that the number of days spent in foster care is important in predicting the exit outcome of a child. However, I cannot claim that staying longer in foster care causes a child to exit one way or another. There can be other unaccounted variables that impact both the length of stay and the exit outcome, such as the child's personality. The fact that the number of days is predictive of outcome exit is not surprising when we consider that the outcomes include adoption and emancipation. Length of stay is correlated with age, and the age of a child is highly predictive whether the child ages out of the foster care system or gets adopted.

Related Work

Studies in social services have long been interested in how the placement experience and personal characteristics of a child affect the child's well being. The main difficulties with the studies are fully capturing the factors that can affect the result, including both the myriad of observable and unobservable variables. Hence, researchers cannot easily recover the causal impact of foster care on the child. For instance, simply comparing the outcomes of children of children who spent time in foster care with those that did not leads to biased estimate. The difference in outcome can be partially or fully explained by the pre-placement experience, such as the extent of the abuse that the child suffered prior to entering. One method to address this issue is to have more extensive data on both the child's placement and pre-placement situations, as done in (Berger, Johnson, Bruch, James, & Rubin, 2009). The most convincing studies are (Doyle, 2007) and (Doyle, 2008), which uses the random assignment of social workers to different cases as instrumental variables. They find that putting children whose cases are on the margin into foster care result in worse child outcome, in terms of health condition and criminal activities.

With the recent passage of Fostering Connection Act of 2008, which increased the benefit to the state child welfare services that the federal government fund, there are longitudanol studies looking at the outcomes of children out of foster care. (Courtney, Charles, & Nathanael J. Okpych, 2014) has collected extensive surveys of the children aged from 16.75 - 17.75 during the year of 2012, and found that the children in the sample had better experience than the national baseline. The

children in the sample were more likely to skip a grade and less likely to drop out of high school.

Data

The data comes from Adoption and Foster Care Analysis and Reporting System (AFCARS). This dataset is a case-level dataset on an annual basis, which means each observation pertains to one child for a given fiscal reporting year.

I restricted the years from 2002 - 2007, and focused on the Los Angeles County. I wanted to abstract away from cross county differences, which include child welfare policies, income level, population composition, and others. I chose the county of LA because the foster care population of LA county makes up over 60% of the foster care population in California. During this period in LA, there were 9968 exits from foster care. Some observations had missing variables. I discarded all the observations that had at least one missing variable, which yielded 9331 observations.

Labels

I focused my attention on 4 main ways in which a child can exit from the foster care system: reunificiation with the family, adoption, emancipation and runaways. Emancipation means that the child has reached the age of 18 and can no longer stay in the foster care system. The breakdown of the exit outcomes is summarized in table 3. As we can see, the sample is very unbalanced. Majority of the children exit by reunifying with their parents, and about a fourth is adopted. I will address this issue of unbalanced classes in the estimation section.

	Reunification	Adoption	Emancipation	Runaway	Total
Freqency	5971	1957	1117	286	9331
Percentage	64 %	21 %	12 %	3%	100%

Table 1
This table breaks down the different forms a child exited from LA county foster care system between 2002 - 2007.

Features (and feature selection)

The AFCARS dataset contains identifiable information of the child. The variables for each observation may include the child's ethnicity, set of disabilities, reasons for removal from original caretakers, and age. There are multiple features for disabilities and reasons for removal, and each is a binary variable with 1 indicating that such disability or removal reason is applicable and 0 otherwise. The variables are not mutually exclusive, which means multiple disabilities or reasons for removal can be applied to a single child. Variables for disability include mental retardation, visual impairment, and physical disability. Reasons for removal may include neglect, physical abuse, sexual abuse, parental drug-abuse, parental alcohol abuse, etc. In total, there are 5 variables that describe the mental, emotional, and physical state of the child when entering foster care, and 15 variables for the reasons of removal.

To reduce the number of features, I performed PCA on the 5 variables for the health state and 15 variables for the reason for removal. I tried to reduce the features to various number of dimensions, but such processing did not yield much improvement in prediction.

I also reduce the number of features by adding up the binary values of certain features into a larger integer value. A child can have more than one of the 5 variables that describe the health of the child can equal one, i.e the child suffers from both mental retardation and visual impairment. Similarly, the reason a child was removed from home may include, neglect and physical abuse. The sum of these binary values may summarize the severity of a child's disability or the severity for the reason for removal. For instance, we might suspect that the situation of a child who was removed for neglect, physical abuse, and drug abuse by the parent is more severe than a child who was removed only for neglect. However, this feature reduction did not improve the prediction performance either. In fact, the performance worsened. This result indicate that the different reasons for removal and dif-

ferent types of disabilities are not interchangeable when prediction the child's exit outcome.

Estimation

I use all the features without reducing them to a lower dimension. I randomly break the 9331 dataset into 70 % training set and 30 % test set.

To address the issue of unbalanced classes, I weigh the penalty of incorrectly classifying a class by the proportion in the training set. I implement using the SVM library in Python, using the option "class_weight."

Since I used sciki-kit learn's svm library, the mathematical formulation is as follows:

$$\min_{w,b,\epsilon} \frac{1}{2} w^T w + \sum_{i}^{n} C_i \epsilon_i$$

$$s.t.$$

$$y_i(w^T \phi(x_i) + b) \geq 1 - \epsilon_i$$

$$\epsilon_i \geq 0 \text{ for } i = 1, ..., n$$

, where C_i corresponds to the penalty assigned to the i^{th} example based on the class weight and n is the number of examples. The python package of support vector classification (SVC) implement an one-versus-one algorithm. The algorithm constructs 4*3/2 = 6 classification coefficients, each for the possible pairing of the 4 labels. When making prediction, an example is predicted using every one of the 6 classification, and the class that receives the most vote is predicted.

I tried 3 different types of kernels: Gaussian, linear, and polynomial. The kernels are parametrized as follows:

$$\text{Gaussian} : \quad K(x, y) = \exp(-\gamma|x - y|), \quad (1)$$

$$\text{Linear} : \quad K(x, y) = < x, y >, \quad (2)$$

$$\text{Polynomial} : \quad K(x, y) = (< x, y > + r)^d, \quad (3)$$

Below is a table that summarizes the accuracy of the training and testing set accuracy for the different kernels: It does not seem that the kernel choice from the three option made a huge difference in the performance of the predictions.

	Gaussian	Polynomial	Linear
Training	76 %	68 %	70 %
Testing	68 %	66 %	67%

Table 2

The parametrization of the kernels are the following: Gaussian - $\gamma = 1$, Polynomial: $d = 5$ and $r = -2$.

I then the confusion matrix to see which categories are most often confused. The results are present in table 3. The columns represent the actual labels and the rows represent the predicted labels.

	Reunification	Adoption	Emancipation	Runaway
Reunification	1718	401	254	79
Adoption	52	152	41	4
Emancipation	29	7	58	3
Runaway	0	1	0	0

Table 3

Confusion Matrix of Testing Set

As seen from the table above, the prediction for reunification is over represented. Even though the actual proportion of reunification is 64 %, that class is predicted 86% of the time. This fact may be driven by how unbalanced the training set is, even with the different weights for the different classes.

I then examine the incremental impact on the testing set accuracy by removing each of the features. The result is summarized below. I only included features that had made at least 0.5 % difference.

As shown by the table, the number of days spent in foster care seems important in determining the exit outcome of the child. For this project, I choose to use class weights rather than discarding examples to balance the training set, because I would have had to discard too much of the data that I had. In the future, I will try to develop an algorithm that randomly selects a subset of examples "emancipation," "reunification," and "adoption."

Features	Change in Accuracy
Mental Retardation	- 0.48 %
Emotionally Disturbed	- 0.5%
Drug Abuse - Child	-0.67 %
Other Medical Condition	-0.5 %
# of days in foster care	-4.2%

Table 4
This table summarize the impact on the accuracy of the testing set prediction when removing a specific feature.

Discussion

One major concern I have of the estimation is the fact that the training set was very unbalanced. Another weakness with the data set is that many of the variables are binary rather than continuous variables. For instance, I simply observe whether a child suffers from mental retardation or visual impairment. However, I do not observe the extent of each of the condition. Hence, we do not know how comparable one variable is to another. Lastly, the fact that the length of stay was most important in predicting the exit outcome may not be very informative. It is correlated with the age of the child, which is correlated with whether a child ages out or gets adopted.

References

Berger, L., Johnson, E., Bruch, S. K., James, S., & Rubin, D. (2009).
Child Development, *80*, 1856âĂŞ1876.

Courtney, M. E., Charles, P., & Nathanael J. Okpych, K. H., Laura Napolitano. (2014). Findings from the california youth transitions to adulthood study (calyouth).

Doyle, J. (2007). Child protection and child outcomes: Measuring the effects of foster care. *The American Economic Review*, *97*.

Doyle, J. (2008). Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care. *Journal of Political Economics*, *116*.