1. What is the problem you want to solve?

Can early studies or surveys of children and their families lead to prediction of key indicators later in life thus paving the way for targeted interventions to improve outcomes for disadvantaged youths? Is there certain demographic information about the parents or children that help us predict what will happen later in life.

1. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?

The Fragile Families Challenge is being sponsored by Princeton University and is bring sociologists and data scientists from around the country to tackle this data. They want to be able to either find ways to intervene with children to improve their futures or discover the limits of the current study and data-set so that improved data can be collected. With my analysis, they will combine my model with others’ to make an improved combination model that will hopefully approach the limits of the data. They also hope to potentially identify new features that would be worth obtaining.

What important fields and information does the data set have?

It seems from early exploration that some of the 1st year demographic information and the most recent economic data are most important. These include:

'cm1ethrace': Race of the mother

'cf1ethrace': Race of the father

'cm1edu': Education level of the mother

'cf1edu': Education level of the father

'cm5povco': A measure of poverty level of the mother

'cm1relf': A measure of mother's relationship with father (married, cohabitating, unknown, etc)

'cm1age': Mother's age when child was born

What are its limitations i.e. what are some questions that you cannot answer with this data set?

The data struggles to determine causality. For instance, do certain types of hardships cause eviction, or is eviction simply another form of hardship and there are further underlying causes of both? The data is also extremely lacking in the categorical predictive features, and so simply logistic regression was not effective.

What kind of cleaning and wrangling did you need to do?

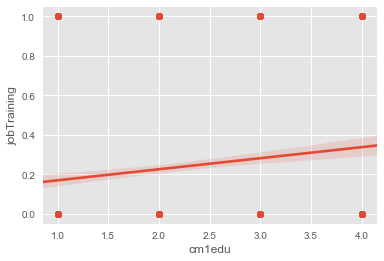
The predictive variables had many missing rows and so those had to be dropped for modeling – thus reducing the amount of available data. The constructed variables had little missing data, so that wasn’t an issue here other than having to find good variables. If I want to include a wider range of features I will have to impute some of the missing data or risk having too little to work with.

Are there other datasets you can find, use and combine with, to answer the questions that matter?

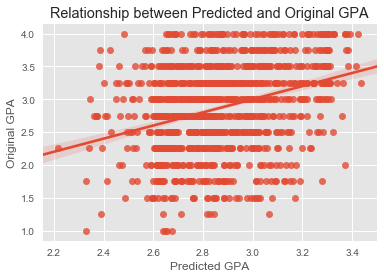
No, but hopefully this project will help understand the limitations of this study and be able to inform future studies on what additional information could be collected.

Any preliminary exploration.

Initial attempts at logistic regression were not successful. Another approach will be needed. This image shows how the no (zero) counts greatly outweigh the yes (1) counts.



The initial linear regression did show some promise, though and did show the general trend.



Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?

I will be using random forest (boosted, too) as well as elastic net to see if I can get an improvement in performance and to better handle the data. Especially with the categorical predictors, another method is needed.