Fragile Families Report

**Findings**

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.

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.

The strength of the data set is also its biggest headache for doing analysis – namely that there is a lot of data. Specifically, there are over 12,000 features. Within these features there are many that are not of significant use in modeling. Many contain a large percentage of missing data, other contain data that is labeled but means that the question was a skip or refused to answer, and others essentially duplicate other columns. Thus, feature selection was a major focus.

In analyzing the data, I focused primarily on what are known as constructed variables. These have been created from the raw data and contain much less missing data. In fact, there were no NaN’s, but there were still some data labeled as “missing” in an intentional way. For this analysis I chose to keep that data as it was labeled and essentially treat it as a new category.

Exploratory Data Analysis was done for several features comparing them to GPA and Job Training. In the end, seven features were chosen for the model. These included age of the mother, race, poverty measure, and education levels.

Initially I performed a simple linear regression for GPA. The results were not very significant with R2 near 0. Improvement was made using a Random Forest Regressor. While R2 only increased to about 0.1, the accuracy was actually around 30%. However, the big break through was to think of GPA as categorical instead of continuous. The data only had GPA broken down by every quarter point, so by creating 13 categories I used a Random Forest Classifier instead and found accuracy to improve to 70%. This was a significant improvement and this route warrants further exploration. The most significant features were the mother’s age when the child was born and the mother’s poverty status when the child was 9 years-old.

Exploration of whether a primary caregiver had undergone some sort of job training was next. As this was a binary yes/no classification, initially I did a logistic regression. The results were horrible. The model simply predicted that everything was a no. The data set has significantly more no entries than yes entries and this gave the model fits. Using Random Forest helped significantly. With that 72% accuracy was obtained. There were still too many false positives and false negatives, but there were at least positives being predicted. This could potentially benefit from additional features. Of note, I did perform normalization of the data, but performance actually decreased so it was left out of the final model.

**Next Steps and Recommendations**

Including more features and performing imputation on the missing data would be the next steps. Perhaps there are raw features that will also be instructive in these measures. It will also be of interest to predict the other 4 categories: grit, material hardship, eviction, and layoff. Future data collection should likely focus on both poverty and mother’s age as those were major features for both categories. While less significant that those two, the third most significant feature was mother and father relationship.

With this in mind, the following recommendations are given:

1. Interventions should focus on mothers who were especially young when the child was born, and who have a poor relationship with the father.
2. Persistent poverty correlated with decreased GPA and a lower likelihood of receiving job training (though which leads to what in that case is difficult to determine). Thus, targeting families that remain in difficult circumstances should be the focus of extra attention for both school and job training opportunities.
3. Future data should be collected on not only was job training done but offered. Additionally, impact data should be collected. After job training, is there economic mobility, or do the interventions increase GPA?