Draft 2 of Capstone Proposal

Analyzing Social Mobility with Data Science

Preface:

Opportunity Insights is a team of researchers and policy analysts work together to analyze new data and create a platform for local stakeholders to make more informed decisions and is based out of Harvard University.

This capstone is heavily based upon two "Empirical Projects" (one and four) from Opportunity Insights' class Using Big Data to Solve Economic and Social Problems taught at Harvard University during the Spring 2019 semester.

The atlas.csv and atlas_training.dta datasets and all other files used are publicly accessible through the class' webpage linked above.

Part 1, telling a data-based story using the Opportunity Atlas

The Opportunity Atlas was publicly released on October 1, 2018, and an accompanying article appeared on the front page of the *New York Times*. The Opportunity Atlas is a freely available interactive mapping tool that traces the roots of outcomes such as poverty and incarceration back to the neighborhoods in which children grew up.

Policymakers, journalists, and the public have begun to explore the Opportunity Atlas, casting new light on the geography of upward mobility in communities across the country. As an example, see Jasmine Garsd's recent analysis for the New York City neighborhood of Brownsville in Brooklyn.

I will use the Opportunity Atlas mapping tool and the underlying data to describe and compare equality in 10 major cities in the United States: (I could cut it down to 5 if I do not have enough time or the analysis is too challenging)

1. New York City
2. San Francisco
3. Los Angeles
4. Chicago
5. Houston
6. Seattle
7. Philadelphia
8. Boston

9. Nashville
10. Denver

This part of the project focuses on the following methods for descriptive data analysis, primarily using the Opportunity Atlas and accompanying atlas.csv dataset:

1. Data Visualization: maps are a powerful way to present descriptive statistics for data with a geographic component.
2. Regression and correlation analysis: using linear regressions and correlation coefficients to quantify the statistical relationship between upward mobility and potential explanatory variables.

Part 2, using Google DataCommons data to predict social mobility

I will use variables from Google DataCommons to predict intergenerational mobility using machine learning methods. The measure of intergenerational mobility that we will focus on is the mean rank of a child whose parents were at the 25th percentile of the national income distribution in each county (kfr*pooled*p25). The goal is to construct the best predictions of this outcome using other variables, an important step in creating forecasts of upward mobility that could be used for future generations before data on their outcomes become available. There are about 5,000 possible covariates available from Google DataCommons! There are 63 predictors in the atlas.csv data set already. Part of the goal is to carefully select at least 10 more predictors from Google DataCommons to use in the prediction algorithm.

This part of the project showcases the following three data science concepts.

1. Data Wrangling and Exploratory Data Analysis (EDA)
2. Prediction using linear regression, decision trees using 10 fold cross-validaton, random forest
3. Out of Sample Validation
4. Data Visualization

Data Wrangling Part 1

1. Find out all the census tracts in each of the major cities
2. Using the atlas.csv file, filter out all the other tracts that are not within those cities.
3. Create a training set and test set. I could also create new training/test sets with only data from the specific cities by merging the current training/test sets provided and then split them again. The benefit to doing it that way is because the provided training/test sets have 121 features compared to the atlas.csv file only having 63

4. Create new datasets for each major city by filtering for them in the atlas.csv file.

Analysis Part 1

1. (To answer this question, read the Opportunity Atlas manuscript)
    1. What period do the data you are analyzing come from?
    2. Are you concerned that the neighborhoods you are studying may have changed for kids now growing up there?
    3. What evidence do Chetty et al. (2018) provide suggesting that such changes are or are not important?
    4. What type of data could you use to test whether your neighborhood has changed in recent years?
2. How does average upward mobility, pooling races and genders, for children with parents at the 25th percentile (kfr pooled*p25*) in each city compare to mean (population-weighted, using count_pooled) upward mobility in the state and in the U.S. overall?
    1. Do kids in those cities have better or worse chances of climbing the income ladder than the average child in America?
    2. How do the cities compare to each other?
3. What is the standard deviation of upward mobility (population-weighted) in each city?
    1. Is it larger or smaller than the standard deviation across tracts in the city's home state?
    2. Across tracts in the country?
    3. How do the cities compare to each other?
4. What can we learn from these comparisons?
5. Now let's turn to downward mobility:
    1. Repeat questions (1) and (2) looking at children who start with parents at the 75th and 100th percentiles.
    2. How do the patterns differ?

Regression and Correlation Analysis

1. Using a linear regression, estimate the relationship between outcomes of children at the 25th and 75th percentile in the cities.
    1. Generate a scatter plot to visualize this regression.
    2. Do areas where children from low-income families do well generally have better outcomes for those from high-income families, too?
    3. We could also compare outcomes of the children at each percentiles in different cities
2. Can I identify any covariates which help explain some of the patterns that have been identified above?
    - Some examples of covariates include housing prices, income inequality, fraction of children with single parents, job density, etc.

    1. For 2 or 3 of these, report estimated correlation coefficients along with their 95% confidence intervals.
3. Formulate a hypothesis for why there is variation in upward mobility for children who grew up in the Census tracts of each city and provide correlational evidence testing that hypothesis.
4. Putting together all the analyses:
    1. What is learned about the determinants of economic opportunity where you grew up?
    2. Identify one or two key lessons or takeaways that you might discuss with a policymaker or journalist if asked about your hometown. Mention any important caveats the conclusions.
        - For example, can we conclude that the variable identified as a key predictor in the question above has a causal effect (i.e., changing it would change upward mobility) based on that analysis? Why or why not?

Data Wrangling Part 2

1. Go to Google DataCommons and select at least 10 county-level variables that might be useful in predicting the statistic that we are using to describe intergenerational mobility which is the variable kfr*pooled*p25.
2. Select and download at least 10 predictors in DataCommons for each of the cities.
    1. First, select a geography and choose predictors. Next, click "Get Code/Data".
    2. Then, click "Bulk Download data." Picking a particular year will generate a .csv file that contains the data for all counties.
        - Note that some data are only available in certain years, so you should pick a year where the variables you want to use are available)
3. Merge the data for each city with its corresponding data set which was created in Part 1.
4. Merge all of the datasets to create a pooled set of just the Census tract for the cities.
5. Split the data into a training and test split.
6. Many of the Google DataCommons variables are counts (e.g., total number of female residents of a county or owner-occupied housing units). Replace these counts with rates (e.g., percent female or fraction of owner-occupied housing units) by dividing by the population and housing variables given in atlas_training.dta.
7. Produce simple summary statistics for the 10 predictors selected from DataCommons and kfr*pooled*p25 in the combined data set for observations that exist in both data sets.

Prediction

1. Run a linear regression of kfr*pooled*p25 on the 10 predictors (converted to rates when appropriate) from Google DataCommons, inspect the results, and comment on what is

found.

2. How well does the linear regression predict krf*pooled*p25 in-sample?
3. Run a linear regression of krf*pooled*p25 on the full predictor set (consisting of the 10 predictors chosen from DataCommons and the 121 predictors included in the training data). Interpret one of the coefficients. Obtain predictions of krf*pooled*p25.
4. Implement a decision tree on the full predictor set using 10 fold cross-validation to select the optimal tree size. What is the first split? Discuss why the first split is often an important predictor or correlate of the outcome.
5. Why do we use cross-validation to select a smaller tree instead of just using as many splits as possible and using a larger tree that would have had a lower prediction error.
6. Implement a random forest with at least 1000 bootstrap samples and obtain predictions.
7. Calculate and compare the mean squared error for the results on 1, 2, and 4 **in -sample**.

Out-of-sample validation

8. Briefly comment on whether or not you think the regression from question 8, question 9 or from question 11 will predict krf*pooled*p25 better **out-of-sample.**
9. Now turn to the test data set. Calculate the mean squared error for the results from 1, 2, and 4 out-of-sample.
10. Which model did the best? Show the in-sample and out-of-sample mean squared error for the models estimated in questions 1, 2, and 4
11. Draw some graphs or maps to visualize the predictions.

Using Super Learner

The Super Learner, a commonly used algorithm in biostatistics, is an ensembling machine learning approach that combines multiple algorithms into a single algorithm and returns a prediction function with the best cross-validated mean squared error.

17. Use the p4_SL.R file to implement to Super Learner algorithm and obtain predictions.
18. Add these new predictions to the answers to questions 7, 9, 10, and 11.

Part 3, How is income linked to social mobility?

- Does household income have a direct effect on a child's social mobility?
  - * NEED HELP HERE
  - Can find an annual income data set by zip code, but the tract*outcomes*early.csv data set (too big too upload to GitHub) has all outcomes in correlation to income and separated by census tract
  - Can formulate a hypothesis test

- Do a time series/arima model on how parental/household income correlates to children's outcome over the years.