Empirical Project 3
Using Google DataCommons to Predict Social Mobility

In this Empirical Project, you 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). Your 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.

The "training" dataset is a 50% random sample of all counties with at least 10,000 residents available from the Opportunity Atlas. You will use predictors from Google DataCommons to predict the variable kfr_pooled_p25 in the *other* half of these data. There are about 5,000 possible covariates available from Google DataCommons! We have included 121 predictors in these data already. Part of the assignment is to carefully select at least 10 more predictors from Google DataCommons to use in your prediction algorithm.

The assignment has three parts:

1. *Data set up*. In the first part, you are asked to select at least 10 predictors from DataCommons. Download these data for all counties using the Bulk Downloads link. Merge these data with the atlas_training.dta data file. Produce descriptive statistics and run a simple linear regression using these combined data.

2. *Prediction challenge*. The second section is about using the training data to construct a prediction algorithm that produces good out-of-sample predictions of krf_pooled_p25.

3. *Out-of-sample validation*. After completing Part 2, you will evaluate your predictions in the test data, which consists of the *other* half of the data. For this part, you will use the atlas_test.dta data file. You will merge your predictions from part 2 with these data and assess the performance of your prediction algorithm.

Instructions

There will be **two deadlines**:

1. **Part 1: SUNDAY MARCH 6TH, 2022 (11:59PM)**
2. **Part 3: SUNDAY MARCH 13TH, 2022 (11:59PM)**

Your submission should include **three files**:

1. A word or pdf document with responses to the questions asked below.
2. A do-file with your STATA code
3. Relevant .dta files (**only for Part 1)**

Part 1: Data set up

1. Go to Google DataCommons and select at least 10 **county-level** variables that you think might be useful in predicting the statistic that we are using to describe intergenerational mobility which is the variable kfr_pooled_p25.

2. Download at least 10 predictors in DataCommons for all counties in the United States. First, select a geography and choose predictors. Then, click "Bulk Download data." This will generate a .csv file that contains the data for all counties.

3. Merge these data with the atlas_training.dta data file.

4. If there are variables with no observations (i.e., all observations are missing) drop such variables.

5. 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 to you in atlas_training.dta.
(Note that Google DataCommons is still under development; although you can draw graphs with per capita figures, only the counts can be downloaded via the Bulk Downloads).

6. Produce simple summary statistics for the 10 predictors you selected from DataCommons and krf_pooled_p25 in the combined data set for observations that exist in both data sets.

7. Run a linear regression of krf_pooled_p25 on the predictors you chose (converted to rates when appropriate) from Google DataCommons, in addition to the predictors already in the training dataset inspect the results, and comment on what you find.

8. How well does your linear regression predict krf_pooled_p25 in-sample? You have to calculate the mean squared error. (Use Table 2)
Submit your answer to these questions, the dofile and the project4.dta file, by 11:59 p.m. on March 6th to receive credit.

Part 2: Prediction Challenge (Not Graded: Done on R)
(The R-code to do this will be given or the TA will create this dataset for you)

9. Run a linear regression of krf_pooled_p25 on the full predictor set (consisting of the 10 predictors you chose from DataCommons and the 121 predictors included in the training data). Obtain predictions of kfr_pooled_p25.

10. Implement a decision tree on the full predictor set using 10 fold cross-validation to select the optimal tree size.

11. Implement a random forest with at least 1000 bootstrap samples and obtain predictions.

12. Calculate the mean squared error for your results.

Part 3: Out-of-sample validation

13. Load in the proj4_results.dta provided to you. Keep only the variables geoid kfr_pooled_p25 test training predictions_ols predictions_tree predictions_forest.

14. Merge the test dataset using the geoid variable.

15. Calculate the mean squared error for predictions_ols predictions_tree predictions_forest out-of-sample. (Hint: Refer Table 2)

16. Which model did the best?

Bonus: Draw some graphs or maps to visualize your predictions.

Submit your answer to these questions and the dofile, by 11:59 p.m. on March 13th to receive credit.

DATA DESCRIPTION, FILE: atlas_training.dta

The data consist of all 2,518 counties with at least 10,000 residents available from the Opportunity Atlas. For n = 1,259 counties in the "test" portion of the data, the outcome variable is set to missing. These observations are a 50% random sample of all counties with at least 10,000 residents available from the Opportunity Atlas. For more details on the construction of the variables included in this data set, please see Chetty, Raj, John Friedman, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. 2018. "The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility." NBER Working Paper No. 25147.

Variable	Definition	Obs.
(1)	(2)	(3)
geoid	County FIPS code	2,518
pop	County Population from DataCommons	2,518
housing	Total number of housing units from Census	2,518
kfr_pooled_p25	Mean percentile rank in the national distribution of household income in 2014-2015 for children with parents at the 25th percentile of the national income distribution (missing for n = 1,259 counties in the test data, non-missing for the other n = 1,259 counties)	1,259
test	1 = Observation is in test data set (outcome variable is missing) 0 = Observation is in training data (outcome variable is non-missing)	2,518
training	1 = Observation is in training data set (outcome variable is non-missing) 0 = Observation is in the test data (outcome variable is missing)	2,518
P_1 through *P_121*	Predictors taken from the Opportunity Insights' county characteristics file and various other sources	2,518

Note: Full list of definitions of *P_1* through *P_121* is posted on the class website.

DATA DESCRIPTION, FILE: atlas_test.dta

Variable	Definition	Obs.
(1)	(2)	(3)
kfr_actual	Actual value for *kfr_pooled_p25* for all 2,518 counties with at least 10,000 residents	2,518
geoid	County FIPS code	2,518

Table 2
Stata Commands

Commands	Description
**clear the workspace* *clear all* **change working directory and open data set* *cd "C:\Users\gbruich\Ec1152\Projects\"*	This code shows how to clear the workspace, change the working directory, and open a Stata data file. To change directories on either a mac or windows PC, you can use the drop down menu in Stata. Go to file -> change working directory -> navigate to the folder where your data is located. The command to change directories will appear; it can then be copied and pasted into your .do file.
import delimited "export.csv", clear	This commands show how to import a .csv file into stata. You can also use the drop down menu.
rename county **	These commands show how to rename variables by removing the county prefix from any variable starting with county.
merge 1:1 geoid using atlas_training.dta, gen(mtrain)	These commands show how to merge the data in current working memory with the training data. The key that connects them is *geoid*. The option *gen(mtrain)* will generate a new variable *mtrain* that marks indicates which observations matched up across the two data sets. See this tutorial for more details.
replace xvar = xvar/pop	This command shows how to replace the variable *xvar* with a rate per person instead of a count.
replace xvar = xvar/housing	This command shows how to replace the variable *xvar* with a rate (fraction of housing units) instead of a count.
save project4.dta, replace	This command saves the data that is currently in the working memory. It will be saved to the working directory (which can be changed as shown above).
gen pred_error = kfr_actual - predictions_forest *gen mse_forest = pred_error^2* *sum mse_forest if test == 1*	This command shows how to report the mean squared prediction error for the test sample. First, we generate prediction errors and squared prediction errors. Then, we summarize this variable for observations in the test sample.