**Lab 7: Predicting Social Mobility using Cross Validation and Random Forests**

[R Labs]

*Methods/concepts: loops, steady states, random forests, cross validation*

**LAB DESCRIPTION**

This is the second labs on prediction policy questions. In this lab, you will predict upward mobility using *decision trees* and *random forests*. The measure of upward mobility that we will focus on is **Statistic 1: Absolute Mobility at the 25th Percentile** in each county (**kfr\_pooled\_pooled\_p25**). For more details on the variables included in these data, see [Table 1](#Table1).

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 121 community characteristics to predict the variable **kfr\_pooled\_pooled\_p25**. The other half of these data has been set aside as a “lock box” data set that you will use to evaluate your models. In the R labs, we will start from starter scripts that can either be run on your computer or on the [FAS On Demand server](https://fasrc.github.io/fas-ondemand-user-guide/).

**QUESTIONS**

1. Primer on for loops in the context of steady states. We will start with a review of the calculation from the Lecture where Professor Chetty introduced the concept of a steady state (Becker and Tomes 1979). This review will also give us an opportunity to walk through loops step by step. Chetty et al. (2020) report the following rank-rank regression pooling all races and genders:

Using the sample code, show that this model predicts convergence in incomes across racial groups. This result is unrealistic because racial disparities have persisted for many generations in the U.S.

However, the model is incorrect: we know from Lab 2 and Lecture that children of different races experience very different rates of upward mobility across generations. In particular, Chetty et al. (2020) report the following rank-rank regression for Black children:

and for Hispanic children:

Use a for loop to find the **steady state prediction** of the model for Black and Hispanic children.

1. Explain briefly how cross-validation helps us avoid the overfit problem.
2. Modify the example code to implement **five-fold cross validation** to choose the depth of a decision tree that uses just two predictors. It is your choice of which two predictors. The predictors are the variables P\_1 through P\_121 in the data, but their real names are included in the data dictionary available in [Table 3](#Table3) below.
   1. Plot the cross-validation pseudo out-of-sample root mean squared prediction error (CV RMSE) versus the depth of the tree.
   2. Using the graph that you produced, what tree depth is optimal?
   3. Now use the full training data set to estimate a tree of the depth you selected in the previous question. Visualize the tree. Which predictors are being used in the first several splits of the tree?
   4. Obtain predictions in the training sample.
3. Explain briefly how random forests improve upon decision trees using (i) bagging and (ii) input randomization.
4. Now implement a random forest with at least 1000 trees (bootstrap samples) using the same two predictors you selected for the decision tree. Obtain predictions in the training sample.
5. Next, implement a random forest with at least 1000 trees (bootstrap samples) using the full predictor set (consisting of the 121 predictors corresponding to variables P\_1 through P\_121 included in the training data). Obtain predictions in the training sample.
6. Random forests typically result in improved accuracy over prediction using a single tree. Unfortunately, however, it can be difficult to interpret the resulting model. Recall from Lab 6 that one of the advantages of decision trees is the attractive and easily interpreted diagram that results.

One can obtain an overall summary of the importance of each predictor in a random forest by measuring how the mean squared error decreases when the predictor is used define tree splits. A large value indicates an important predictor.

Using the random forest from the previous question, which variables are the most important predictors using this metric? Refer to the data dictionary in [Table 3](#Table3) below to determine what these variables measure.

1. Calculate and compare the root mean squared prediction error for your three models in the **training sample**. Which model does the best?
2. Now turn to the lock box data set **atlas\_lockbox.dta**. These data contain a variable called **kfr\_actual** which is the “truth:” the actual value of **kfr\_pooled\_pooled\_p25** for all the counties in the sample, including the 50% of the data in the lock box sample. Calculate predictions from your models and use **kfr\_actual** to calculate the root mean squared prediction error for the test sample. Which model did the best?
3. For this Lab, please submit the following:
4. Your final Python Jupyter notebook (you can insert your answers to the above questions directly into the notebook) or .R script file to Gradescope
5. A single PDF document with the answers and graphs submitted to Gradescope. You can print the Python Jupyter notebook as a PDF using the file drop down menu.
6. There will also be a Google form that is projected to the screen in Lab

**DATA DESCRIPTION, FILE: atlas\_training.dta [training data]**

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](https://opportunityinsights.org/wp-content/uploads/2018/10/atlas_paper.pdf).

TABLE 1

Training Data

|  |  |  |
| --- | --- | --- |
| **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\_pooled\_p25* | Statistic 1: Absolute Mobility at the 25th Percentile  **(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 in [Table 3](#Table3) on the next page.

**DATA DESCRIPTION, FILE: atlas\_lockbox.dta [Lock box data]**

TABLE 2

Lock Box Data

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition** | **Obs.** |
| (1) | (2) | (3) |
| *kfr\_actual* | Actual value for *kfr\_pooled\_pooled\_p25* for all 2,518 counties with at least 10,000 residents | 2,518 |
| *geoid* | County FIPS code | 2,518 |

TABLE 3

Complete List of All Predictor Variables in Training Data

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | Variable | Description | Obs. |
|  | (1) | (2) | (3) |
|  |  |  |  |
| 1 | *geoid* | County FIPS code | 2,518 |
| 2 | *pop* | County Population from DataCommons | 2,518 |
| 3 | *housing* | Total number of housing units from Census | 2,518 |
| 4 | *kfr\_pooled\_pooled\_p25* | Statistic 1 Absolute Mobility at the 25th Percentile | 1,259 |
| 5 | *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 |
| 6 | *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 |
| 7 | *P\_1* | *Bankruptcies per 1000 adults in 2008* | 2,518 |
| 8 | *P\_2* | Bankruptcies per 1000 adults in 2009 | 2,518 |
| 9 | *P\_3* | Bankruptcies per 1000 adults in 2010 | 2,518 |
| 10 | *P\_4* | Bankruptcies per 1000 adults in 2011 | 2,518 |
| 11 | *P\_5* | Bankruptcies per 1000 adults in 2012 | 2,518 |
| 12 | *P\_6* | Bankruptcies per 1000 adults in 2013 | 2,518 |
| 13 | *P\_7* | Bankruptcies per 1000 adults in 2014 | 2,518 |
| 14 | *P\_8* | Bankruptcies per 1000 adults in 2015 | 2,518 |
| 15 | *P\_9* | Bankruptcies per 1000 adults in 2016 | 2,518 |
| 16 | *P\_10* | % of Individuals Earning < 138% of the FPL without Insurance in 2013 | 2,518 |
| 17 | *P\_11* | % of Individuals Earning 138%-400% of the FPL without Insurance in 2013 | 2,518 |
| 18 | *P\_12* | Total Violent and Property Crimes Rate | 2,518 |
| 19 | *P\_13* | Total Violent Crimes Rate: Murder Rate | 2,518 |
| 20 | *P\_14* | Total Violent Crimes Rate: Rape Rate | 2,518 |
| 21 | *P\_15* | Total Violent Crimes Rate: Robbery Rate | 2,518 |
| 22 | *P\_16* | Total Violent Crimes Rate: Aggravated Assault Rate | 2,518 |
| 23 | *P\_17* | Total Property Crimes Rate | 2,518 |
| 24 | *P\_18* | Total Property Crimes Rate: Burglary Rate | 2,518 |
| 25 | *P\_19* | Total Property Crimes Rate: Larceny Rate | 2,518 |
| 26 | *P\_20* | Total Property Crimes Rate: Motor Vehicle Theft Rate | 2,518 |
| 27 | *P\_21* | Total Violent and Property Crime Arrests Rate | 2,518 |
| 28 | *P\_22* | Total Violent and Property Crime Arrests Rate: Violent Crime Arrests Rate | 2,518 |
| 29 | *P\_23* | Total Violent and Property Crime Arrests Rate: Property Crime Arrests Rate | 2,518 |
| 30 | *P\_24* | Mean Household Income 2000 | 2,518 |
| 31 | *P\_25* | Average Commute Time of Working Adults in 2000 | 2,518 |
| 32 | *P\_26* | Fraction of Residents w/ a College Degree or More in 2000 | 2,518 |
| 33 | *P\_27* | Fraction of Residents w/ a College Degree or More in 2006-2010 ACS | 2,518 |
| 33 | *P\_28* | Share of Population Born Outside the U.S. in 2006-2010 ACS | 2,518 |
| 34 | *P\_29* | Median Household Income in 2016 | 2,518 |
| 35 | *P\_30* | Median Household Income in 1990 | 2,518 |
| 36 | *P\_31* | Share Below Poverty Line 2006-2010 ACS | 2,518 |
| 37 | *P\_32* | Share Below Poverty Line 2000 | 2,518 |
| 38 | *P\_33* | Share Below Poverty Line 1990 | 2,518 |
| 39 | *P\_34* | Share black 2010 | 2,518 |
| 40 | *P\_35* | Share hisp 2010 | 2,518 |
| 41 | *P\_36* | Share asian 2010 | 2,518 |
| 42 | *P\_37* | Share black 2000 | 2,518 |
| 43 | *P\_38* | Share white 2000 | 2,518 |
| 44 | *P\_39* | Share hisp 2000 | 2,518 |
| 45 | *P\_40* | Share asian 2000 | 2,518 |
| 46 | *P\_41* | Average School District Level Standardized Test Scores in 3rd Grade in 2013 | 2,518 |
| 47 | *P\_42* | Average Rent for Two-Bedroom Apartment in 2015 | 2,518 |
| 48 | *P\_43* | Share of Single-Headed Households with Children 2006-2010 ACS | 2,518 |
| 49 | *P\_44* | Share of Single-Headed Households with Children 1990 | 2,518 |
| 50 | *P\_45* | Share of Single-Headed Households with Children 2000 | 2,518 |
| 51 | *P\_46* | Share of Working Adults w/ Commute Time of 15 Minutes Or Less in 2006-2010 ACS | 2,518 |
| 52 | *P\_47* | Employment Rate 2000 | 2,518 |
| 53 | *P\_48* | Census Form Rate Return Rate 2010 | 2,518 |
| 54 | *P\_49* | Log wage growth for HS Grad., 2005-2014 | 2,518 |
| 55 | *P\_50* | Share of People who are not white 2010 | 2,518 |
| 56 | *P\_51* | Population Density (per square mile) in 2010 | 2,518 |
| 57 | *P\_52* | Population Density (per square mile) in 2000 | 2,518 |
| 58 | *P\_53* | Average Annual Job Growth Rate 2004-2013 | 2,518 |
| 59 | *P\_54* | Job Density (in square miles) in 2013 | 2,518 |
| 60 | *P\_55* | Physically Unhealthy Days per Month (Persons 18 Years and Over) | 2,518 |
| 61 | *P\_56* | Mentally Unhealthy Days per Month (Persons 18 Years and Over) | 2,518 |
| 62 | *P\_57* | Percent of Adults That Report Fair or Poor Health (Persons 18 Years and Over) | 2,518 |
| 63 | *P\_58* | Percent of Low Birthweight Births (<2.5kg) | 2,518 |
| 64 | *P\_59* | Primary Care Physicians (PCP) Rate per 100,000 Population | 2,518 |
| 65 | *P\_60* | Mental Health Providers (MHP) Rate per 100,000 Population | 2,518 |
| 66 | *P\_61* | Dentists Rate per 100,000 Population | 2,518 |
| 67 | *P\_62* | Health Care Costs Price-adjusted Medicare Reimbursements | 2,518 |
| 68 | *P\_63* | Percent of Persons Without Insurance (Population Under 19 Years, 2013 est.) | 2,518 |
| 69 | *P\_64* | Percent of Persons Without Insurance (Population 18 to 64 Years, 2013 est.) | 2,518 |
| 70 | *P\_65* | Percent of Persons Without Insurance (Population Under 65 Years, 2013 est.) | 2,518 |
| 71 | *P\_66* | Premature Age-adjusted Mortality Rate per 100,000 Population | 2,518 |
| 72 | *P\_67* | Drug Poisoning Mortality Rate per 100,000 Population | 2,518 |
| 73 | *P\_68* | Percent Diabetics (Adults) | 2,518 |
| 74 | *P\_69* | Percent of Diabetic Medicare Enrollees Receiving Hba1c Test | 2,518 |
| 75 | *P\_70* | Diabetic Medicare Enrollees (Out of Total Medicare Enrolles) | 2,518 |
| 76 | *P\_71* | Teen Births Rate per 100,000 Population (Females 15 to 19 Years) | 2,518 |
| 77 | *P\_72* | Chlamydia Cases Rate per 100,000 Population | 2,518 |
| 78 | *P\_73* | HIV Prevalence Rate per 100,000 Population | 2,518 |
| 79 | *P\_74* | Percent Current Smokers (Persons 18 Years and Over) | 2,518 |
| 80 | *P\_75* | Percent Drinking Adults (Persons 18 Years and Over) | 2,518 |
| 81 | *P\_76* | Percent of Persons with Limited Access to Healthy Foods | 2,518 |
| 82 | *P\_77* | Percent of Persons with Access to Exercise Opportunities | 2,518 |
| 83 | *P\_78* | Percent Obese Persons (20 Years and Over) | 2,518 |
| 84 | *P\_79* | Percent Percent Physically Inactive Persons (20 Years and Over) | 2,518 |
| 85 | *P\_80* | Percent of Children Eligible for Free Lunch (Persons < 18 Years) | 2,518 |
| 86 | *P\_81* | Food Environment Index | 2,518 |
| 87 | *P\_82* | % Total: Evangelical Protestant | 2,518 |
| 88 | *P\_83* | % Total: Mainline Protestant | 2,518 |
| 89 | *P\_84* | % Total: Historically Black Protestant | 2,518 |
| 90 | *P\_85* | % Total: Roman Catholic | 2,518 |
| 91 | *P\_86* | % Total: Jewish Congregations | 2,518 |
| 92 | *P\_87* | % Total: Latter-day Saint (Mormon) | 2,518 |
| 93 | *P\_88* | % Total: Islamic | 2,518 |
| 94 | *P\_89* | % Total: Hindu | 2,518 |
| 95 | *P\_90* | % Total: Buddhist | 2,518 |
| 96 | *P\_91* | % Total: Orthodox Christian | 2,518 |
| 97 | *P\_92* | % Total: Jehovah's Witnesses | 2,518 |
| 98 | *P\_93* | % Total: Other | 2,518 |
| 99 | *P\_94* | % Total: Evangelical Protestant Member Count | 2,518 |
| 100 | *P\_95* | % Total: Mainline Protestant Member Count | 2,518 |
| 101 | *P\_96* | % Total: Historically Black Protestant Member Count | 2,518 |
| 102 | *P\_97* | % Total: Roman Catholic Member Count | 2,518 |
| 103 | *P\_98* | % Total: Jewish Member Count | 2,518 |
| 104 | *P\_99* | % Total: Latter-day Saint (Mormon) Member Count | 2,518 |
| 105 | *P\_100* | % Total: Islamic Member Count | 2,518 |
| 106 | *P\_101* | % Total: Hindu Member Count | 2,518 |
| 107 | *P\_102* | % Total: Buddhist Member Count | 2,518 |
| 108 | *P\_103* | % Total: Orthodox Christian Member Count | 2,518 |
| 109 | *P\_104* | % Total: Jehovah's Witnesses Member Count | 2,518 |
| 110 | *P\_105* | % Total: Other Member Count | 2,518 |
| 111 | *P\_106* | % Total Evangelical Protestant: Advent Christian Church | 2,518 |
| 112 | *P\_107* | % Total Evangelical Protestant: Adventists - Other | 2,518 |
| 113 | *P\_108* | % Total Evangelical Protestant: Church of God General Conference | 2,518 |
| 114 | *P\_109* | % Total Evangelical Protestant: Seventh Day Adventists | 2,518 |
| 115 | *P\_110* | % Total Evangelical Protestant: Seventh Day Church of God | 2,518 |
| 116 | *P\_111* | % Total Evangelical Protestant: American Baptist Association | 2,518 |
| 117 | *P\_112* | % Total Evangelical Protestant: Baptist General Conference | 2,518 |
| 118 | *P\_113* | % Total Evangelical Protestant: Baptist - Other | 2,518 |
| 119 | *P\_114* | % Total Evangelical Protestant: Baptist Bible Fellowship | 2,518 |
| 120 | *P\_115* | % Total Evangelical Protestant: Baptist Missionary Association of America | 2,518 |
| 121 | *P\_116* | % Total Evangelical Protestant: Cooperative Baptist Fellowship | 2,518 |
| 122 | *P\_117* | % Total Evangelical Protestant: Independent Baptist Churches | 2,518 |
| 123 | *P\_118* | % Total Evangelical Protestant: Conservative Baptist Association | 2,518 |
| 124 | *P\_119* | % Total Evangelical Protestant: Free Will Baptists | 2,518 |
| 125 | *P\_120* | % Total Evangelical Protestant: General Assoc. of Regular Baptists | 2,518 |
| 126 | *P\_121* | % Total Evangelical Protestant: Assoc. of General Baptists | 2,518 |