Final Project Analysis Plan

Chloe Dinh and Matyas Chlebovsky
ISOM 330

Dataset

Community and Crime Dataset
(https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime)

Abstract

Communities within the United States. The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR.

Research question

Can we use the various independent variables present in the dataset to predict crime (Per Capita Violent Crimes), and based on that, recommend policies to help prevent/lowering crime rates?

Variables

Response variable:

Per Capita Violent Crime – calculated using population and the sum of crime variables considered violent crimes in the U.S.: murder, rape, robbery, and assault (UCI Machine Learning Repository Website)

Predictors:

- State (numeric): state numeric code for each community
- Population (numeric): total population of the community
- Householdsize (numeric): average household size
- Racepctblack (numeric): percentage of population that is African American
- RacePctWhite (numeric): percentage of population that is Caucasian
- RacePctAsian (numeric): percentage of population that is of Asian heritage
- RacePctHisp (numeric): percentage of population that is of Hispanic heritage
- AgePct12t21 (numeric): percentage of population that is 12-21 in age
- AgePct12t29 (numeric): percentage of population that is 12-29 in age
- AgePct16t24 (numeric): percentage of population that is 16-24 in age

- AgePct65ip (numeric): percentage of population that is 65 and over in age
- numbUrban (numeric): number of people living in areas classified as urban
- pctUrban (numeric): percentage of people living in areas classified as urban
- medIncome: median household income
- pctWWage: percentage of households with wage or salary income in 1989
- pctWFarmSelf (numeric): percentage of households with farm or self-employment income in 1989
- pctWInvInc (numeric): percentage of households with investment / rent income in 1989pctWInvInc: percentage of households with investment / rent income in 1989
- pctWSocSec (numeric): percentage of households with social security income in 1989
- pctWPubAsst (numeric): percentage of households with public assistance income in 1989
- pctWRetire (numeric): percentage of households with retirement income in 1989
- medFamInc (numeric): median family income (differs from household income for non-family households)
- perCapInc (numeric): per capita income
- whitePerCap (numeric): per capita income for Caucasians
- blackPerCap (numeric): per capita income for african americans
- indianPerCap (numeric): per capita income for native americans
- AsianPerCap: per capita income for people with asian heritage (numeric - decimal)
- OtherPerCap (numeric): per capita income for people with 'other' heritage
- HispPerCap (numeric): per capita income for people with hispanic heritage
- NumUnderPov (numeric): number of people under the poverty level
- PctPopUnderPov (numeric): percentage of people under the poverty level
- PctLess9thGrade (numeric): percentage of people 25 and over with less than a 9th grade education
- PctNotHSGrad (numeric): percentage of people 25 and over that are not high school graduates
- PctBSorMore (numeric): percentage of people 25 and over with a bachelors degree or higher education
- PctUnemployed (numeric): percentage of people 16 and over, in the labor force, and unemployed
- PctEmploy (numeric): percentage of people 16 and over who are employed
- PctEmplManu (numeric): percentage of people 16 and over who are employed in manufacturing
- PctEmplProfServ (numeric): percentage of people 16 and over who are employed in professional services
- PctOccupManu (numeric): percentage of people 16 and over who are employed in manufacturing
- PctOccupMgmtProf (numeric): percentage of people 16 and over who are employed in management or professional occupations
- MalePctDivorce (numeric): percentage of males who are divorced
- MalePctNevMarr (numeric): percentage of males who have never married
- FemalePctDiv (numeric): percentage of females who are divorced
- TotalPctDiv (numeric): percentage of population who are divorced

- PersPerFam (numeric): mean number of people per family
- PctFam2Par (numeric): percentage of families (with kids) that are headed by two parents
- PctKids2Par (numeric): percentage of kids in family housing with two parents
- PctYoungKids2Par (numeric): percent of kids 4 and under in two parent households
- PctTeen2Par (numeric): percent of kids age 12-17 in two parent households
- PctWorkMomYoungKids (numeric): percentage of moms of kids 6 and under in labor force
- PctWorkMom (numeric): percentage of moms of kids under 18 in labor force
- NumIlleg (numeric): number of kids born to never married
- PctIlleg (numeric): percentage of kids born to never married
- NumImmig (numeric): total number of people known to be foreign born
- PctImmigRecent (numeric): percentage of immigrants who immigrated within last 3 years
- PctImmigRec5 (numeric): percentage of immigrants who immigrated within last 5 years
- PctImmigRec8 (numeric): percentage of immigrants who immigrated within last 8 years
- PctImmigRec10 (numeric): percentage of immigrants who immigrated within last 10
- PctRecentImmig (numeric): percent of population who have immigrated within the last 3 years
- PctRecImmig5 (numeric): percent of population who have immigrated within the last 5 years
- PctRecImmig8 (numeric): percent of population who have immigrated within the last 8 years
- PctRecImmig10 (numeric): percent of population who have immigrated within the last 10 years
- PctSpeakEnglOnly (numeric): percent of people who speak only
- PctNotSpeakEnglWell (numeric): percent of people who do not speak English well
- PctLargHouseFam (numeric): percent of family households that are large (6 or more)
- PctLargHouseOccup (numeric): percent of all occupied households that are large (6 or more people)
- PersPerOccupHous (numeric): mean persons per household
- PersPerOwnOccHous (numeric): mean persons per owner occupied household
- PersPerRentOccHous (numeric): mean persons per rental household
- PctPersOwnOccup (numeric): percent of people in owner occupied households
- PctPersDenseHous (numeric): percent of persons in dense housing (more than 1 person per room)
- PctHousLess3BR (numeric): percent of housing units with less than 3 bedrooms
- MedNumBR (numeric): median number of bedrooms
- HousVacant (numeric): number of vacant households
- PctHousOccup (numeric): percent of housing occupied
- PctHousOwnOcc (numeric): percent of households owner occupied
- PctVacantBoarded (numeric): percent of vacant housing that is boarded up
- PctVacMore6Mos (numeric): percent of vacant housing that has been vacant more than 6 months
- MedYrHousBuilt (numeric): median year housing units built
- PctHousNoPhone (numeric): percent of occupied housing units without phone (in 1990, this was rare!)

- PctWOFullPlumb (numeric): percent of housing without complete plumbing facilities
- OwnOccLowQuart (numeric): owner occupied housing - lower quartile value
- OwnOccMedVal (numeric): owner occupied housing - median value
- OwnOccHiQuart (numeric): owner occupied housing - upper quartile value
- RentLowQ (numeric): rental housing - lower quartile rent
- RentMedian (numeric): rental housing - median rent (Census variable H32B from file STF1A)
- RentHighQ (numeric): rental housing - upper quartile rent
- MedRent (numeric): median gross rent (Census variable H43A from file STF3A - includes utilities)
- MedRentPctHousInc (numeric): median gross rent as a percentage of household income
- MedOwnCostPctInc (numeric): median owners cost as a percentage of household income - for owners with a mortgage
- MedOwnCostPctIncNoMtg (numeric): median owners cost as a percentage of household income - for owners without a mortgage
- NumInShelters (numeric): number of people in homeless shelters
- NumStreet (numeric): number of homeless people counted in the street
- PctForeignBorn (numeric): percent of people foreign born
- PctBornSameState (numeric): percent of people born in the same state as currently living
- PctSameHouse85 (numeric): percent of people living in the same house as in 1985 (5 years before)
- PctSameCity85 (numeric): percent of people living in the same city as in 1985 (5 years before)
- PctSameState85 (numeric): percent of people living in the same state as in 1985 (5 years before)
- LandArea (numeric): land area in square miles
- PopDens (numeric): population density in persons per square mile
- PctUsePubTrans (numeric): percent of people using public transit for commuting
- LemasPctOfficDrugUn (numeric): percent of officers assigned to drug units

Data Understanding and Preprocessing:

We will be using Python to clean, process, and perform exploratory data analysis.

Processing

The list of predictors above does not include a couple of variables in the original datasets that have many NA values. We will exclude those variables from the final analysis

The original dataset does not include the header column. We will add this header column back into the data before we run the analysis.

Most numeric values from the dataset have already been normalized into the decimal range 0.00-1.00 using an Unsupervised, equal-interval binning method (UCI Machine Learning Repository Website). We will not go through this processing step.

The state codes are in numeric values. We will convert this into its abbreviation form for easier interpretation. The corresponding codes are found on the Census website: https://www.census.gov/geo/reference/ansi_statetables.html.

We will divide the data into testing and training sets (70-30 ratio).

Candidate Model

We will be using regression to predict Per Capita Violent Crime. Depending on the preliminary results of the linear model, we might try to boost it by using LDA and/or Ridge regression.