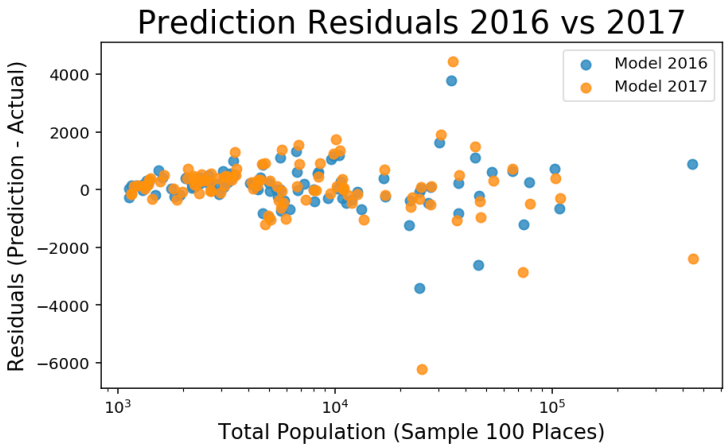
Quantitative Population Prediction by Place (USA)

Winston Robson | Galvanize Data Science Immersive g88 (San Francisco, Spring 2019) | https://bit.ly/2Z5IONb

Purpose

This project aims to utilize machine learning on combined Census and American Community Survey datasets to predict the future population of any place in the United States.

Outcome



- 2016
  - Model performed 25.4% better than Baseline

Place	Model Pred	Baseline Pred	Actual Pop
San Francisco	862,737	864,025	850,282
New York	8,491,355	8,584,487	8,461,961
New Orleans	335,259	396,463	382,922
Houston	2,297,084	2,300,705	2,240,582
Bentonville, AR	39,239	43,605	42,499
Sidney, NE	6,879	6,890	6,875
Pleasanton, CA	74,550	78,530	77,046

- 2017
  - Model performed 49.8% better than Baseline

Place	Model Pred	Baseline Pred	Actual Pop
San Francisco	872,139	887,287	864,263
New York	8,549,441	8,742,231	8,560,072
New Orleans	330,641	416,188	388,182
Houston	2,344,416	2,382,704	2,267,336
Bentonville, AR	39,984	46,842	44,601
Sidney, NE	6,900	6,950	6,830
Pleasanton,CA	73,549	82,051	79,341

Process

1. Exploratory Data Analysis
   a. Examined large number of Geographic filters on Total Population
      i. E.g. Place, 5-digit Zip, County
   b. Determined Place to be most usable
      i. Counties were too ranged in Number of Measurements, some since 1790 and having 20+ w/ others < 10
      ii. 5-digit Zip measurements were initialized too recently, no multi-decade historical data

2. Combined Place Census 1970-2010 Total Population data with each Place ACS 5-year Estimate 2011-2015 of Total Population (Age/Sex)
   a. Forgetting places that do not coexist across datasets (would not accurately measure)

3. Defined Baseline as assuming Continued Trend from Place's population change

4. Engineered Generalized Additive Time-Series Model using Facebook's Prophet to forecast Total Population

5. Randomly sampled 100 places Coexisting across ACS years 2011-2017 and 2+ Census 1970-2010
   a. Allows interpretation on per year and decades-long basis
   b. Did not consider places with Total Population less than 1,000; unrealistic predictions

6. Fit Model on each sample place

7. Measured Model Outcomes v. Baseline Assumptions
   a. Compared to Actual 2016 & 2017 Populations
   b. Model consistently outperforms when taking a sample (multiple Places), but in specific instances does come up short

Tech Used

Project on Github Winston Robson LinkedIn 479.268.0047 / 925.400.8172 warobson@gmail.com