

On random sample of 100 Places

E.g.

San Francisco, California

Bentonville, Arkansas

New Orleans, Louisiana

West Fairlee, Virginia

Model is 15.9% more accurate than Baseline for 2016

Model is 6.2% more accurate than Baseline for 2017

Model is 11.04% more accurate than Baseline overall

1-pager

Quantitative Population Prediction by Place in the United States and Puerto Rico

Background:

Property prices, like all things, fluctuate. What drives these fluctuations? Is there one indicator that could be used to anticipate these fluctuations?

Data:

US Census measurements of Total Population per Place in the United States and Puerto Rico 1970-2010

American Community Survey 5-year Estimates of Total Population per Place in the United States and Puerto Rico by Age & Sex 2011

American Community Survey 5-year Estimates of Total Population per Place in the United States and Puerto Rico by Age & Sex 2012

American Community Survey 5-year Estimates of Total Population per Place in the United States and Puerto Rico by Age & Sex 2013

American Community Survey 5-year Estimates of Total Population per Place in the United States and Puerto Rico by Age & Sex 2014

American Community Survey 5-year Estimates of Total Population per Place in the United States and Puerto Rico by Age & Sex 2015

American Community Survey 5-year Estimates of Total Population per Place in the United States and Puerto Rico by Age & Sex 2016

American Community Survey 5-year Estimates of Total Population per Place in the United States and Puerto Rico by Age & Sex 2017

Challenges:

The American Community Survey is a very new thing.

The United States of America has changed geographically numerous of times since the first US Census.

Measurement levels in the US Census have changed numerous times since the first US Census in {YEAR}.

5-digit Zip Code since 2000

Place since 1970

County by State since 1790

Data Scientist Words:

* Exploratory Data Analysis
  + Clustering
    - 5-digit Zip Code
  + Timeseries
    - 5-digit ZIp Code
    - Not enough data
  + County
    - Too much data
    - Points per County ranged
      * Some since 1790 (23 points)
      * Large amount with far less
  + Place
    - Reasonable amount of data
    - Large amount since 1970 (5 points)
      * Issue
        + 2010 was a long time ago and predicting 2010 drops max points to 4
      * Solution
        + Predict 2020

Issue

Test would be on 2010, same issue as before

Solution

Predict next year; test will then be on this year

Issue

Data is by decade

Solution

Introduce ACS Survey for yearly information

* + Issue
    - Completely different information
  + Solution
    - Clean; merge; test on 2016 and 2017, providing 5 decade and 5 yearly measurements
  + Model
    - Timeseries
      * Facebook’s fbprophet
        + Prophet model
        + Forecast total population by place
  + Clustering
    - sklearn KMeans
      * Cluster by Age & Sex breakdown of 5-digit Zip Code

Would like to expand to qualitative predictions

> Specific

> Working Population per Employer

> Student v. Teacher Ratio

> Tenure vs Age

> General

> Housing

> Industry

> Taxes

> Both at scale

Tech used: <https://github.com/facebook/prophet>