An Early Warning System for Gentrification in California

Aakash Pattabi | Department of Economics | apattabi@stanford.edu

Introduction and Problem

Despite California's high incomes, the astronomical cost of housing has driven more Californians into poverty than in any other state.

Understanding the drivers of gentrification and accurately forecasting when neighborhoods will undergo change will be instrumental as policymakers design responses to California's housing crisis.

This research applies machine learning techniques to forecasting gentrification in Census Tracts in California.

Goals

- Characterize CA's housing markets with public data
- Accurately forecast gentrification in Tracts
- Understand the drivers of gentrification through feature selection and regularization

Data

Data comes from *American FactFinder (AFF)*, a public repository of local-, state-, and national-level Census data collated by the United States Census Bureau. It is at the **Census Tract** level.

Census Tracts are hyper-local geographic bounding boxes containing ~4,000 people. They are generally invariant in scope over time.

AFF releases inter-censal surveys with housing market data such as:

- Renter-occupied vs. owner-occupied unit counts
- Educational attainment of renters vs. owners
- Race, ethnicity, age of Census Tract residents
- Employment by industry and job tenure

This project's data were assembled from tables `S2502`; `S2503`; `B25085`; and `DP03` in *AFF*, comprising ~150 features.

Features and Responses

Responses:

- Long-term Δ in monthly housing costs
- Long term Δ in income distribution, measured by shift in **Hellinger distance** of income from start-year to end-year vs. a baseline (see **Figure 1**). Defined over discrete distributions *P(X), Q(X)* as:

$$\Delta_{Hell} = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} \left(\sqrt{P(X = x_i)} - \sqrt{Q(X = x_i)} \right)^2}$$

Engineered features:

- First order spatial lag in housing cost and income distribution shift
- Local Moran's I-statistic of spatial clustering (see **Figure 2)**

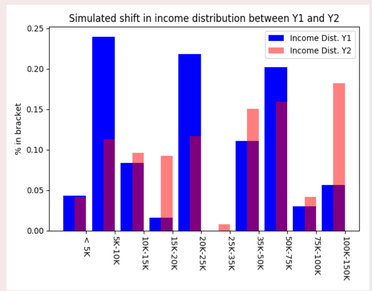

Figure 1: This simulated income distribution becomes more affluent (and less tri-modal) from Y1 to Y2. We argue this is indicative of gentrification

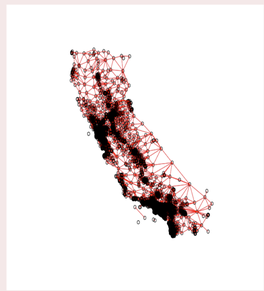

Figure 2: California's Census Tracts were modeled as an unweighted, undirected graph to engineer features based on the features of adjacent Tracts

Tobler's First Law of Geography: Everything is related to everything else, but nearer things moreso than further ones.

Models and Results

Models:

Random Forest, Gini Loss

$$G(E) = 1 - \sum_{i=1}^{k} P(i \mid E)^2$$

Trees split on random subsets of features to min. Gini impurity in leaves

L1-Penalized Logistic Regression

Let $P(x)$ be logistic. LASSO solves:

$$L(\theta) = -\sum_{i=1}^{m} y_i P(x_i) + (1 - y_i)(1 - P(x_i))$$

$$\theta = argmin_\theta \, L(\theta) + \frac{1}{C} \sum_{i=1}^{n} |\theta_i|$$

XGBoost

Ensemble of shallow trees where *subsequently grown predictors depend on previous ones*

Minimizes logistic loss $L(\theta)$ (defined left) where $P(x)$ is given by proportion of trees voting positive in a given iteration

Ensemble

Ensemble predicts the class that received the most votes of the random forest model, penalized logit, and XGBoost

Results:

- Models trained on *difference* in feature values from 2010 to 2011, responses calculated from 2012 to 2016
- Hyperparameters for all models tuned by grid search (random forest: no. trees, split subset size; LASSO: regularization coefficient; XGBoost: learning rate, stump depth, and regularization coefficient)

Response: Δ in monthly cost of housing over time				
Model n_{train} = 7,262 n_{test} = 397	**Test Accuracy**	**Precision**	**Recall**	**No Info Rate**
Random Forest	0.62	0.64	0.69	0.53
L1-Penalized Logit	0.58	0.59	0.70	0.53
XGBoost	0.64	0.65	0.69	0.53
Ensemble	0.63	0.63	0.71	0.53

Table 1: XGBoost and the ensemble outperform other classifiers and improve meaningfully over simply predicting majority class in sample

Response: Δ in income distribution over time				
Model n_{train} = 7,262 n_{test} = 397	**Test Accuracy**	**Precision**	**Recall**	**No Info Rate**
Random Forest	0.58	0.58	0.85	0.59
L1-Penalized Logit	0.55	0.58	0.85	0.59
XGBoost	0.53	0.59	0.70	0.59
Ensemble	0.56	0.58	0.86	0.59

Table 2: No classifier beats no info rate for income distribution response. High recall, low precision suggest "trigger-happy" positive labelling…

Discussion

- Ex-ante balanced-ness of the classes was surprising; suggests gentrification is spatial (some countervailing economic force is ensuring costs don't rise uniformly)
- Non-parametric estimators (Random Forest, XGBoost) outperformed logit; likely due to near-inability to overfit at no cost of bias
- Grid search shows large accuracy gains from regularizing (LASSO: $C = 0.005$, XGB: $\lambda = 25$)
- Signal was much stronger in cost than income distribution response; not surprising given they were almost uncorrelated ($\rho = 0.06$)

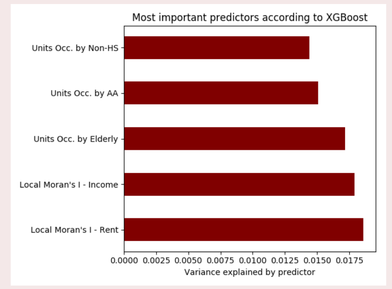

Figure 3: Spatial features and features correlated with (lack of) affluence are influential

- Feature importance ranking validated engineering of spatial features (see **Figure 3**)

Future Work and References

- Construct adjacency matrix weighted by e.g. inter-Tract centroid distance to encode "decaying" influence into engineered features
- Simplify income distribution response by collapsing buckets to reduce noise, increase signal

Veronica Guerrieri, Daniel Hartley, and Erik Hurst. "Endogenous Gentrification and Housing Price Dynamics". In: NBER Working Paper Series (2010).

Ken Steif. "Predicting gentrification using longitudinal census data". In: Urban Spatial (2016).

Miriam Zuk. "Regional Early Warning System for Displacement". In: US Department of Housing and Urban Development (2015).

Liana Fox. "The Supplemental Poverty Measure: 2017". In: United States Census Bureau: Economics and Statistics Administration (2018).