A Proximity-Based Early Warning System for Gentrification in California

Aakash Pattabi
Department of Economics
{apattabi@stanford.edu}

December 13, 2018

I. Introduction

With the recent failure of Senate Bill (SB) 827 in California, pressure is higher than ever on state politicians to better understand and respond to the increasing unaffordability of California's urban centers. Designed to issue more housing construction permits in high-opportunity areas, SB 827 was ironically crippled by its failure to explicitly acknowledge the possible gentrification externalities of new housing construction. Because of the astronomical (and increasing) cost of housing, more Californians live in poverty than in any other state when cost of living is accounted for [6]. It is imperative that California's policymakers articulate smart housing policies that do not lock out access to the state's economic engines to the neediest Californians.

One tool that academics use to design thoughtful housing policy is the *gentrification early warning system* [4]. Such systems are frameworks for using state and local data to describe emergent gentrification at a hyper-local level. Previously, [14] analyzed Bay Area data and conducted nine in-depth case studies in Bay Area communities to develop a gentrification classification scheme labelling Census Tracts from "Not Losing Low-Income Households" to "Advanced Gentrification." Unfortunately, conducting focused ethnographic research will not always be possible within policymakers' budgetary and time constraints. While prior machine learning work in this area is sparse, [10] used stepwise discriminant analysis to characterize gentrifying tracts in isolation using demographic and economic features, absent any spatial data or contextual focus. [12] showed significant accuracy gains using off-the-shelf methods incorporating spatial features, but this work was limited to forecasting home prices over time, which may not perfectly correlate with gentrification, especially in California where affordable housing and rent caps are widespread. We extend prior work by:

i. Using California-wide Census data to classify emergent gentrification and to understand the leading indicators of gentrification through feature selection;

ii. And modelling the state's housing market as an interconnected network to test an economic theory of how gentrification spreads.

Specifically, we use machine learning techniques – primarily non-parametric models such as Random Forests and Gradient Boosting – to ascertain the leading indicators of gentrification at the Census Tract level in California. We formulate the problem as binary classification over a five-year time horizon, using custom-designed responses to proxy for whether gentrification was observed in a community over the prediction period.

II. Data: Responses and Features

We source data from American FactFinder (*AFF*), a public information tool produced by the United States Census [2]. We focus on "Census Tracts," local geographic bounding boxes that house on average 4000 people [3]. Using Tract-level data from 2010-2016 from *AFF*, we construct two responses that indicate whether gentrification occurred in a Census Tract.

Prior research describes gentrification in terms of either rising costs of living or displacement of the poor, as income distributions shift towards affluence [14]. To model the first as a response, we use the inter-year, intra-tract difference in the median monthly cost of housing for all residents:

$$y_i = \text{Median Cost}_{i,t'} - \text{Median Cost}_{i,t}$$

Because gentrification occurs over a long time horizon, we split the feature set around a pivot year of 2012; we compute the responses using the data from years 2012-2016 with the data from 2010 and 2011 used as features (in the above formulation, $t = 2012$ and $t' = 2016$). Splitting the data to forecast gentrification over a long time horizon comports with previous research; [12] uses decennial Census estimates. We chose the pivot year by evaluating the performance of the models on an independent validation set for each pivot in $\{2012, 2013, 2014, 2015\}$.

To model the second response, we use an imputed measurement of the inter-year, intra-tract change in the income distribution of the tract (see Table 1).

Less than $5,000
$5,000-$9,999
$10,000-$14,999
$15,000-$19,999
$20,000-$24,999
$25,000-$34,999
$35,000-$49,999
$50,000-$74,999
$75,000-$99,999
$99,999-$149,999
$150,000 or More

Table 1: The Census discretizes income reporting into bins that are more granular towards the lower end of the income scale.

To compute the inter-censal change in the income distribution, we use the *Hellinger distance* measurement of the distance between two distributions. Over two discrete distributions $P(X), Q(X)$ with the same support, the Hellinger distance is:

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

For each tract, we compute the Hellinger distance between the observed income distribution and a baseline in which all residents are perfectly affluent with probability 1. Tracts with low Hellinger distances tend to be high-income; tracts with high Hellinger distances tend to be low-income. Finally, we compute the response by taking the differences of these Hellinger distances for each tract between 2012 (the pivot year) and 2016. A tract that becomes more affluent (gentrifies) from 2012 to 2016 has a negative difference, and vice versa for a tract that becomes more low-income. We rescale the responses so that they are bounded between 0 and 100 and positive differences signal gentrification.

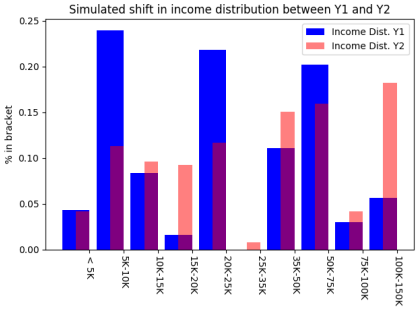

Figure 1: In this Census Tract, the income distribution skews towards affluence (and becomes less trimodal) between Year 1 and Year 2. This shift indicates that gentrification occurred.

Finally, we relabel each response 1 (gentrification occurred) or 0 (gentrification did not occur) for both the monthly cost of housing and income distribution shift responses.

We characterize each Census Tract using a vector of roughly 150 features assembled from tables `S2502`, `S2503`, `B25085`, and `DP03` in *AFF*. These include Tracts' demographic and economic characteristics, such as: employment by industry; ethnic and racial composition; level education; and more.

Additionally, we engineer four features based on the theory of spatial equilibrium proposed in prior work on endogenous gentrification [8]. This theory posits that gentrifying Tracts are highly influenced by the gentrification patterns in their near neighbors. For each Census Tract, we calculate the first order spatial lag and the local Moran's I-Statistic of spatial clustering for both responses during the observation period. We construct these features by modelling California as an unweighted, undirected network with nodes being Census Tracts and edges occurring between Tracts that are adjacent (by queen's contiguity [13]). The two first order spatial lag features describe the change in the average cost of living and in the income distribution between 2010 and 2011 for Census Tracts surrounding each given Tract. This lag is computed as follows:

$$\text{lag}_i = \frac{\sum_{j:i \leftrightarrow j} y_j}{\sum_j \mathbb{1}\{i \leftrightarrow j\}}$$

Here, y_j denotes each response computed between the pre-pivot years, 2010 and 2011. Likewise, for each response, we compute the Local Moran's I-Statistic, a measure of spatial clustering [12]. Economic theory suggests that spatial randomness in an area's housing

market indicates that the market is in disequilibrium – an indicator of gentrification. Conversely, spatial homogeneity indicates the area is in equilibrium, with a low probability of gentrifying. We compute this feature for each response as:

$$\text{Moran's I}_i = \frac{Z_i}{\left(\frac{\sum_j Z_j^2}{n}\right)} \sum_{j:i \leftrightarrow j} Z_j$$

Where Z_k is the deviation of the response of interest from the mean across all n Tracts in the training sample (computed between 2010 and 2011, the observation period).

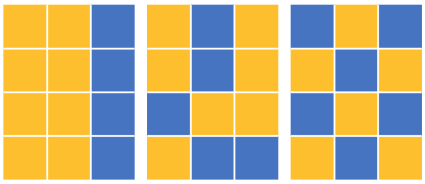

Figure 2: From left to right, these panels display high spatial autocorrelation (clustering); minimal spatial autocorrelation (randomness that tends to indicate market disequilibrium); and spatial anticorrelation.

We do not use time-invariant features describing the geography of the Census Tracts. These ought not add much explanatory power to a model that forecasts gentrification by time. Likewise, we do not add network-topological features from e.g. [7] as Census Tracts are modified or added with extreme rarity [3].

Overall, the data consist of 8,056 observations for each of California's Census Tracts (with one dropped due to missing data). Surprisingly, *a priori* we observed the classes to be roughly balanced for both responses, suggesting that there still exist pockets of affordability in the state. We split the data into a training set comprising 90% (7,262) of the observations and validation and test sets comprising 5% (397) respectively.

III. Methods

We applied four machine learning methods to each classification problem (defining gentrification as the change in monthly cost and as the shift in income distribution). We used a Random Forest classifier; a Gradient Boosting model (XGBoost); an ℓ_1-penalized logistic regression; and an ensemble approach that classified Census Tracts according to a majority vote of the aforementioned three models.

Random Forests are a variant of bagged decision trees; a Random Forest classifier grows a substantial number of independent classification trees each of which minimizes the Gini impurity of its leaf nodes through recursive binary splitting [1]. The Gini impurity of node E given k classes is:

$$G(E) = 1 - \sum_{i=1}^{k} Pr.\{i|E\}^2$$

Gini impurity measures how often a randomly chosen observation in the node would be mislabelled if it were assigned a random label according to the distribution of responses in the node. As classification trees grown on the same set of bootstrapped data tend to be highly correlated, the Random Forest algorithm decorrelates the trees by constraining each split in each tree to be on a random subsample of features in the feature space.

Gradient Boosting is an ensemble technique using classification trees in which trees are grown sequentially (as opposed to simultaneously in Random Forests). Later trees are grown to minimize the errors made by their predecessors. Each subsequent tree "learns" from the mistakes made earlier in training. XGBoost, a popular implementation of Gradient Boosting which enables regularization of the trees, minimizes the loss function:

$$\mathcal{L}(\phi) = \sum_i \ell(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

Where $\hat{y}_i$ are the predicted class; each f_k is a decision tree; and $\Omega(\cdot)$ is a regularization function of the number of leaves in each tree and the weights of those leaves [5]. We used logistic loss as the loss function $\ell(\cdot)$.

For the Random Forest estimator, we tuned n, the number of trees and p, the number of features in the random split set at every split. For the XGBoost estimator, we tuned the learning rate α, the tree depth d on each tree, and the regularization parameter λ.

Our final unitary model was the only parametric estimator – ℓ_1-penalized logistic regression, commonly known as the LASSO. The LASSO estimator is a variation on linear regression that logit-transforms the responses to estimate:

$$\log\left(\frac{Pr.\{y_i = 1|x_i\}}{1 - Pr.\{y_i = 1|x_i\}}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \epsilon$$

Logistic regression models $Pr.\{y_i = 1\}$ as logistic in the features [11]. This estimator imposes a penalty

in the objective function on the size of the parameters β in absolute value. The parameters are:

$$\hat{\beta} = \text{argmin}_{\beta}\mathcal{L}(\beta) + \frac{1}{C}\sum_{i=1}^{n}|\beta_i|$$

Where $\mathcal{L}(\beta)$ is the logistic loss function. Because the LASSO penalizes parameter coefficients in absolute value, it implicitly performs feature selection as features with little predictive power have their parameter coefficients driven to zero. For the LASSO estimator, we tuned the regularization parameter C.

We tuned all hyperparameters via two-stage grid search. First, we drew test hyperparameters uniformly from a representative interval around the model implementations' default parameters in [11]. For example, we initially searched random forest sizes $n \in \{25, 50, 100, 150, 250, 500\}$. Second, we narrowed the grid search to focus on tested hyperparameters around the values that maximized accuracy on an independent validation set in the first stage of the search (see Table 2).

Model	Parameter	Value
Random Forest	n	125
Random Forest	p	18
XGBoost	α	0.25
XGBoost	d	4
XGBoost	λ	25
Logistic Reg.	C	0.005

Table 2: The grid-search values of λ for XGBoost and C for ℓ_1-penalized logistic regression reveal that regularization greatly impacted model performance in classifying Tracts according to the change in the monthly cost of housing. This is likely due to the large feature dimensionality.

The high value of λ and low value of C found by grid search on the validation set suggest that models that perform poorly may be vulnerable to overfitting, especially given the high feature dimensionality.

IV. Discussion

We evaluated each classifier on each of the two responses using accuracy, precision, and recall. While accuracy measures the proportion of test set class assignments that match the true labels, precision and recall provide granular insight into classification errors. Recall quantifies the proportion of positive classes (instantiations of gentrification) that were correctly captured by the classifier; precision quantifies the prediction accuracy solely among the samples that were predicted to be in the positive class.

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

Figure 3: XGBoost and the ensemble model performed best on the change in monthly housing cost response, with a 10% accuracy improvement over the no information classifier.

Precision is commonly used when the cost of false positives is high – such as, when there may be resources wasted in a misdirected policy response. Recall is commonly used when the cost of false negatives is high – such as when families are being displaced. While no one metric dominates in importance in this domain, precision and recall illuminate why the performance of all classifiers on the task of classifying Tracts according to their change in income distribution during the prediction period was so poor.

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

Figure 4: No model outperformed the no information classifier on the income distribution shift response.

All four classifiers outperformed the no information classifier in predicting whether a Tract would gentrify as defined by a rise in the monthly cost of housing (see Figure 3). XGBoost, the ensemble model, and the Random Forest estimators outperformed the no information baseline substantially – by roughly 10 percentage points. Furthermore, all three outperformed the parametric logistic regression, suggesting either some implicit hierarchical structure to the problem or simply that the logit model's inherent bias limited its accuracy.

By contrast, no model outperformed the no information classifier in predicting whether a Tract would

gentrify based on its income distribution. This is not surprising, given how uncorrelated these responses were, with $\rho = 0.06$. The high recalls and relatively low precisions reported by the Random Forest, logit model, and voting classifier suggest a plausible explanation: that all three were overly "trigger-happy" in labelling Tracts as positive instantiations of the response, leading to high counts of true positive labelings (and few false negatives – boosting recall) as well as high counts of false positive labelings (damping precision). The confusion matrix for the Random Forest estimator – the best model on this problem – indicates that the estimator guessed "positive" 86% of the time, an overwhelming majority given that the classes were balanced in the training and test sets (see Table 3). Examining the mislabelled Tracts in greater detail suggests that we may attribute these estimators' noisy performance to the fine granularity of the response. Some Tracts that saw only superficial income distribution changes over the prediction period were particularly susceptible to mislabeling, perhaps because their features were highly similar to Tracts further away spatially that underwent gentrification.

	Pred. 0	Pred. 1
True 0	21	144
True 1	36	201

Table 3: The confusion matrix for the Random Forest on the income distribution response reveals that the classifier predicted positive 86% of the time.

To understand the leading indicators of gentrification, we examine the most important features for the classifiers on the housing cost problem.

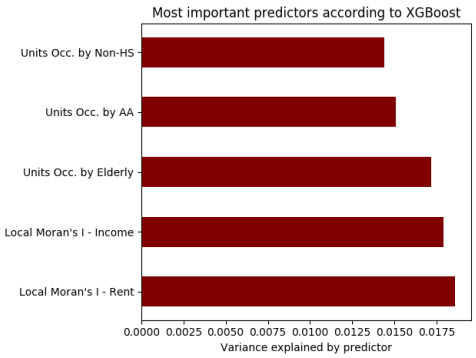

Figure 5: Engineered features of spatial clustering were XGBoost's most important predictors, lending credibility to the theory of endogenous gentrification.

The Local Moran's I-Statistics for income distribution shift and change in the monthly cost of housing were the two most important features extracted from XGBoost, the best-performing model on the independent test set. This lends credence to the theory that gentrification occurs when housing markets are in disequilibrium, indicated by high spatial randomness in their features. That the three next most important predictors quantify the number of elderly people; African Americans; and non-high-school graduates living in each Census Tract is intuitive as well. In California, these groups tend to earn below the median wage [9] and tend to cluster in areas where the cost of living is low (e.g. in the case of the elderly, in retirement communities). These areas tend to be particularly vulnerable to gentrification as residents have few recourses when wealthy urbanites are willing to pay exorbitant prices to move out of even more exorbitantly priced cities.

V. Conclusion

In this research, we develop a classifier to predict whether gentrification will occur in a California Census Tract with 65% accuracy. We defined gentrification as an increase in the inflation-adjusted monthly cost of housing and observed experimentally that other definitions – such as ones based on localities' income distributions – yielded noisy results using public data. Non-parametric ensemble models such as Random Forests and XGBoost outperformed parametric models, which may have overfit the training data. Furthermore, engineered features describing the spatial characteristics of each Census Tract proved most consequential, lending credence to the theory that housing markets in spatial disequilibrium precede gentrification.

Further work might refine the spatially-engineered features by e.g. weighting the network adjacency matrix so that the i, jth entry denotes inverse inter-centroid distance instead of adjacency. Alternatively, further work might focus on better defining gentrification by quantifying displacement of families or collapsing the bins of the income distribution response to increase the signal in the data. Finally, causal work could ascertain the drivers of gentrification as opposed to simply leading indicators. Accurately forecasting gentrification continues to be a pressing problem for California policymakers.

VI. Code

All code written for this project can be found here.

References

[1] Leo Breiman. "Random Forests". In: *Machine Learning* (2001), pp. 5–32.

[2] United States Census Bureau. *American FactFinder*. https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml. Accessed: November 10, 2018.

[3] *Census Tracts*. https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf.

[4] Karen Chapple and Miriam Zuk. "Forewarned: The Use of Neighborhood Early Warning Systems for Gentrification and Displacement". In: *Cityscape* (2016), pp. 109–130.

[5] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Knowledge Discovery in Databases, 2016* (2016).

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

[7] Aditya Grover and Jure Leskovec. "node2vec: Scalable Feature Learning for Networks". In: *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 855–864.

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

[9] Joel Kotkin. *The Hollowing-Out of the California Dream*. https://www.city-journal.org/html/california-economy-16076.html. Accessed: December 11, 2018. 2018.

[10] Han Li. "Modeling gentrification on the census tract level in Chicago from 1990-2000". In: *The University of Toledo Digital Repository: Theses and Dissertations* (2012).

[11] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

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

[13] Fahui Wang. *Quantitative Methods and Socio-Economic Applications in GIS*. 2014.

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