PREDICT GENTRIFICATION

CPLN 5080 Public Policy Analysis | Neve Zhang, Viva Wan, Yaohan Xu

INTRODUCTION

Chicago, the capital of the Midwest and home to over two million people, has evolved from its industrial past[1]. However, this progress has also introduced challenges, such as post-industrial transformation and enduring segregation and disinvestment affecting African American communities[2]. Today, these historic burdens set the stage for reinvestment and gentrification, where affluent groups increasingly settle in upgraded low-income neighborhoods, often resulting in displacement[3].

To help the city identify areas at risk for gentrification and allocate funds accordingly, Marie Antoinette Predictions has developed a predictive model for gentrification incidents in Chicago. The model uses binomial logistic regression to predict whether a census tract is gentrified, primarily based on environmental factors at the census tract level.

The analysis indicates that gentrification in Chicago predominantly occurs in areas where minority communities still represent a significant portion of the population but experience rising income levels and educational attainment. Our model predicts highly effectively in 2015, especially on disadvantaged areas. We are confident that this model will be a valuable tool for the city, enabling targeted policy interventions and efficient allocation of resources.

METHODOLOGY

Our binomial logistic model development follows a multi-tiered approach that includes tract labeling, predictor gathering, and predictor selection.

Tract Labeling

We labeled gentrified census tracts using K-means clustering[4] on demographic and socioeconomic data from 2015 and 2020, including changes from 2010-2015 and 2015-2020. The most representative gentrification pattern in Chicago (Cluster 3 in Table 1) shows significant increases in median household income (over $8,000), high educational attainment (over 10%), and the white population share, along with decreases in Black and Hispanic population shares. These labeled tracts provide the target variable for model development and validation.

Predictor Gathering

Gentrification is often linked to local environmental features and amenities, indicating neighborhood transformation and urban development before demographic and socioeconomic shifts[5]. Therefore, we incorporated environmental factors for both the current situation and the changes over the past five years as predictors using data from 2010, 2015, and 2020[6]. This approach also avoids redundancy and multicollinearity with the demographic and socioeconomic data used in the K-means clustering phase.

Table 1_Results of K-means Clustering (K=5)

cluster	hh.inc	ch.hh.inc	pct.bach	ch.pct.bach	ptc.white	ch.ptc.white	ptc.black	ch.ptc.black	ptc.hispanic	ch.ptc.hispanic
1	34,955	-1,004	18.0%	+1.9%	3.4%	+0.5%	86.7%	-3.3%	7.1%	+2.2%
2	46,836	+244	13.3%	+1.7%	12.8%	-1.7%	6.1%	-0.6%	77.0%	+1.8%
3	62,373	+8,142	47.7%	+11.1%	45.2%	+8.3%	11.6%	-2.0%	30.5%	-7.8%
4	66,179	-923	43.5%	+1.0%	55.6%	-6.8%	9.9%	+0.6%	20.1%	+4.5%
5	116,411	+16,639	75.1%	+5.5%	73.3%	-0.3%	5.9%	-0.6%	10.2%	-0.3%

[1] http://www.encyclopedia.chicagohistory.org/pages/409.html

[2] https://interactive.wttw.com/dusable-to-obama/the-great-migration

[3] https://www.enterprisecommunity.org/sites/default/files/2021-07/Gentrification%20White%20Paper10-9-Final_1.pdf, page 1

[3] https://www.enterprisecommunity.org/sites/default/files/2021-07/Gentrification%20White%20Paper10-9-Final_1.pdf, page 1

[4] K-means clustering, an unsupervised machine learning method, optimizes subgrouping of tracts and minimizes judgment errors.

[5] Chapple, K., & Zuk, M. (2016). Forewarned: The Use of Neighborhood Early Warning Systems for Gentrification and Displacement. Cityscape, 18(3), 109–130. http://www.jstor.org/stable/26328275

[6] Refer to Technical Addendum “3. Predictor Gathering”

Predictor Selection

We developed a binomial logistic model using 2015 predictors and divided the dataset in half for training and testing. To select the final predictors, we logged right-skewed predictors, removed highly correlated predictors, and excluded those that did not contribute to the overall model performance[7]. After finalizing the predictors, we chose the threshold that balanced both specificity and sensitivity[8].

MODEL RESULT

Result Interpretation

The model correctly predicts the status of 73% of all testing tracts in 2015 and reveals associations between gentrification and environmental factors. It shows negative associations between crime counts or increasing crime rates and gentrification, and positive associations between home seeker trends, such as lower vacancy rates and higher in-migration rates, and gentrification. The model also suggests that tracts in better-developed areas, indicated by higher Starbucks density or proximity to transit stations, are more resistant to intensive upgrading and gentrification. (Table2)

Cross Validation for 2015

After spatial cross-validation across Chicago's neighborhoods, we found that the model performs best at predicting gentrification in areas like Logan Square, Garfield Park, Lake View, and Pilsen. However, these neighborhoods with a high true gentrification prediction rate are also prone to have low correct prediction rate on non-gentrified areas. While the model captures non-gentrified areas particularly well, it underestimates several gentrifying neighborhoods such as Portage Park and Goose Island.(Figure1)

Table 2_Model Summary and Data Source

Varibale	Estimate (S.E.)	Data Source
(Intercept)	-1.148*(0.539)	
crime	-0.004**(0.001)	Chicago Data Portal
ch.crime	-0.004*(0.002)	Chicago Data Portal
ch.crime.den	0.001(0.002)	Chicago Data Portal
vacant.rate	-0.035(0.022)	ACS
pct.mig.1yr	0.031*(0.015)	ACS
log.star.den	-0.016(0.074)	SDRDL[9]
sub.nn	-5.0E-05(3.3E-05)	Chicago Data Portal

Significant Level: 0 '***' 0.001 '**' <0.01 '*' 0.05 '.' 0.1' '

According to cross-validation on income level, education level, and race distribution, the model predicts disadvantaged census tracts[10] better than others (Figure 2). It shows higher accuracy and higher correct prediction rate for both gentrified and non-gentrified tracts in those area than others. In other words, the model can better predict the risks of gentrification for disadvantaged tracts compared to wealthier or white-dominated tracts. Notably, it predicts 100% correctly for Hispanic-dominant census tracts experiencing gentrification. However, it is important to note that the model is prone to misclassifying areas with higher income and educational attainment as gentrified, potentially making it difficult to distinguish true gentrification from false instances.

Validation on 2020

After using the 2020 dataset to validate the model, it maintains a relatively high accuracy rate of 64%, despite a slight decrease from 2015. The spatial cross-validations for 2020 show a similar pattern to that in 2015 (Figure 3). Although the model performs worse at distinguishing non-gentrified tracts except in South Chicago, it predicts true gentrifications better in Chicago's north and west ends, such as Lincoln Square and Austin. This signals a new trend of gentrification expanding further north and west. Howeverm northern and western Chicago continues to receive higher misclassification rates as gentrified. In terms of different census tracts considering income, education, and predominant race, the model performs slightly worse than in 2015 for disadvantaged tracts overall (Figure 4). However, it still shows a higher accuracy rate over 80% for disadvantaged tracts compared to others.

RECOMMENDATIONS

We hereby recommend the City select our model for the following strengths:

Key Environment-Conscious Patterns: The model identifies and leverages key associations between tract-level gentrification and environmental factors such as safety, vacancy rates, and Starbucks density. This approach provides environment-based strategies to identify gentrification, avoiding community profiling.

[7] Refer to Technical Addendum “5.2. Variable Selection”

[8] Threshold = 0.18, Sensitivity = 0.73, Specificity = 0.73

[9] San Diego Regional Data Library Data Repository

[10] Disadvantaged census tracts are defined as those with the lowest income levels, lowest education levels, predominant non-white populations, or predominant Hispanic populations.

Locally-Based Data-Driven Method: Our product utilizes a comprehensive data-driven approach, tapping into city data assets like crime records. As these datasets are frequently updated and locally sourced, the model is well-suited for Chicago. Although some predictors are currently retrieved from the census, we are willing to support the City in identifying suitable internal datasets.

Cross-Time Applicability: The model has shown better predictive abilities than a random chance for both years tested. It has correctly identified different areas of gentrification in both years, revealing a spatial pattern of expansion.

Additionally, acknowledging several limitations, we recommend implementing the model with care and anticipate collaboration with the City to address the following issues:

Machine Biases: The model may misclassify areas with higher income and educational attainment as gentrified, potentially leading to extra expenditure for deeper investigation to avoid unnecessary intervention.

Ensuring Model Responsiveness Towards Minority Communities: The model's accuracy in predicting true gentrification among minority-dominant tracts has decreased between 2015 and 2020, raising equity concerns if gentrification instances are under-predicted.

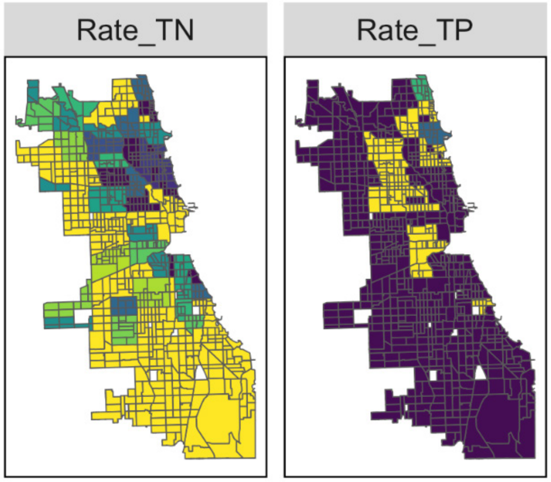

Figure 1_Spatial Cross Validation for 2015

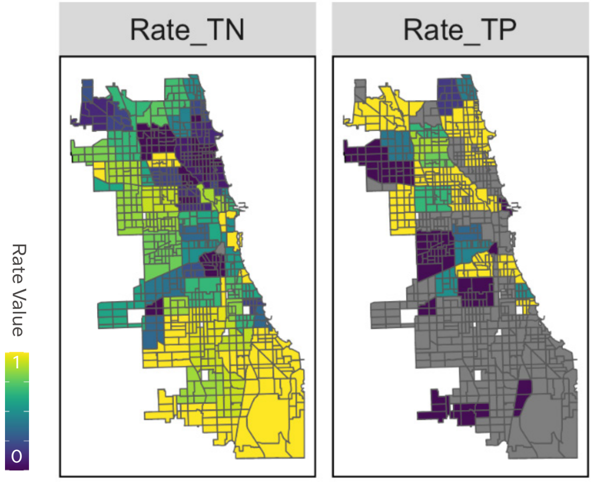

Figure 3_Spatial Cross Validation for 2020

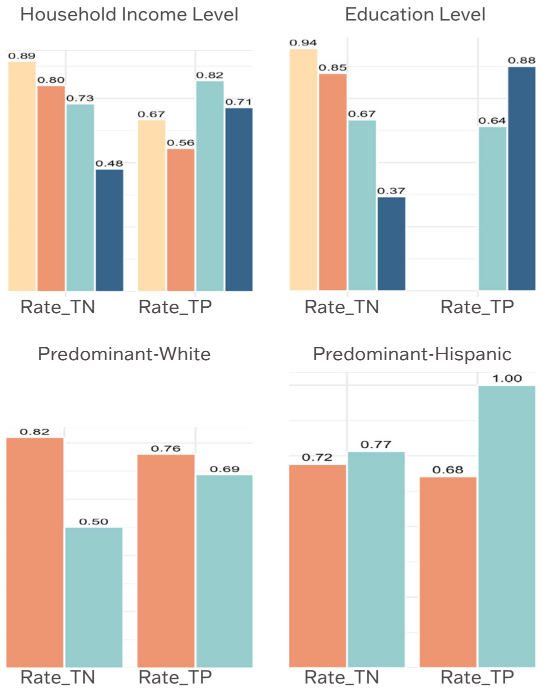

Figure 3_Cross Validation of Socioeconomics for 2015

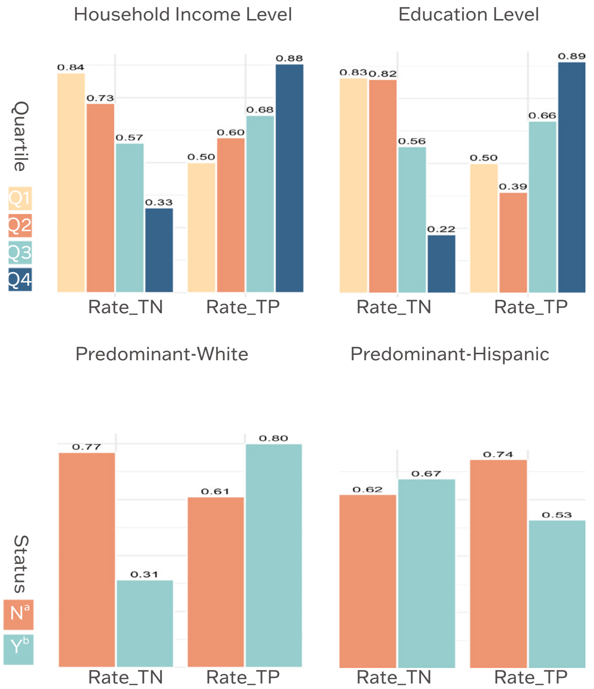

Figure 4_Cross Validation Socioeconomics for 2020

[a] "N" indicates census tracts that are not predominantly White or not predominantly Hispanic.

[b] "Y" indicates census tracts that are predominantly White or predominantly Hispanic.