MUSA508 Public Policy Analytics

Final Report

Gentrification Predictive Model

Jiahang Li, Xiayuanshan Gao, George Chen

Github: https://github.com/XiaysG/PPA_Final

Introduction

Predicting and managing urban gentrification is increasingly crucial for city planners and policymakers aiming to balance economic growth with community sustainability. This report introduces a predictive model for gentrification, focusing on its application on identifying areas at risk within Los Angeles City. The project leverages spatial and census data from 2015 to 2020 for the City of Los Angeles, integrating key metrics such as property values, demographic shifts, and local amenities to construct a robust model of urban transformation.

Methodology

Measuring Gentrification

In our study, we adopted a refined approach to measuring gentrification, drawing upon the methodologies described in prior research, particularly from the National Community Reinvestment Coalition (NCRC) guidelines. Our primary aim was to identify census tracts that underwent significant socio-economic changes indicative of gentrification. To this end, we set criteria based on changes observed from 2015 to 2020, utilizing the American Community Survey (ACS) five-year data sets.

The criteria for defining a tract as gentrified were based on several key indicators:

1. Population Eligibility: Originally, tracts with a population greater than 500 were considered to be eligible tracts for further analysis, but in order to capture all information from the census tract data, we decided to use all tracts instead of the tracts with a population greater than 500 based on the 2014. Eligible tracts would also fit the following criteria in order to be included for further analysis: Median Home Value < 40th percentile and Median Household income < 40th percentile.
2. Socioeconomic Indicators: We assessed changes in median home values, median household income, and the percentage of residents with a college education. A tract was considered at risk of gentrification if it experienced:

- An increase in median home value and median household income greater than the 40th percentile citywide rather than the 60th percentile as outlined in the original NCRC model. This adjustment was made to capture more subtle yet significant shifts that may not reach the higher threshold but still represent meaningful change.
- An increase in the proportion of college-educated residents, also above the 40th percentile, reflects educational gentrification.

3. Metropolitan level Comparison: All data was loaded, calculated, and filtered in the Metropolitan level scale instead of trimming to the city area at the beginning and operating in the city scale, which would allow the gentrification identification to be more accurate, reflecting the gentrification trend in the ambient tracts.

Tracts based on 2020 census information meeting these criterias were classified as 'gentrified' (1) and all others as 'non-gentrified' (0). This methodological approach helps pinpoint the tracts most affected by gentrification based on significant socio-economic transformations. By lowering the percentile threshold for income and home value increases, we include areas undergoing earlier stages of gentrification, which are crucial for timely policy interventions.

Spatial and Non-Spatial Data Collection and Feature Engineering

Our project tests a combination of spatial and non-spatial data to construct a comprehensive predictive model of gentrification in Los Angeles. The spatial data encompasses geographical boundaries of census tracts, locations of amenities such as grocery stores, restaurants, educational facilities, and transit stations. Non-spatial data comprises demographic and socioeconomic variables derived from the American Community Survey (2015-2020), which include population density, age distribution, ethnic composition, income levels, education levels, housing characteristics, and migration patterns.

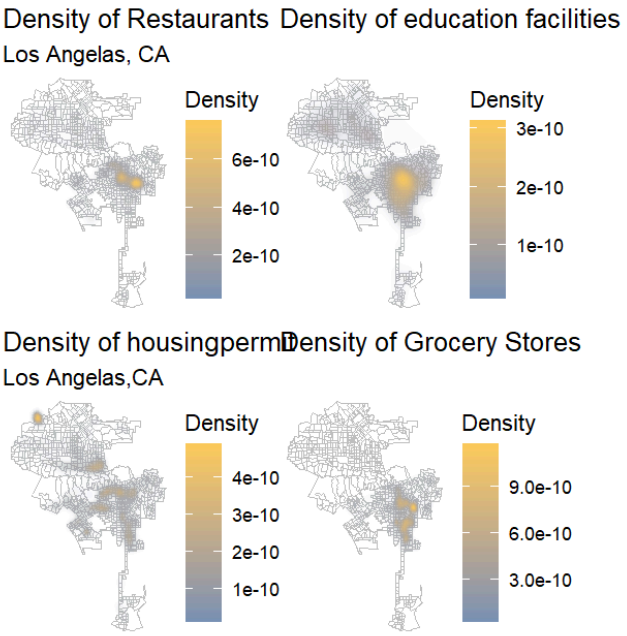
1.Census Data:

As current studies mentioned the impact of historical racial segregation (Hwang & Sampson, 2014) and migration (Hwang et al., 2015) on gentrification, we give attention to white and minority household ownership proportions and migration rates. We also consider more demographic variables referencing the study of DeVylder et al. (2019), including gender, age group, annual household income, and education level. We refined the raw census data to calculate percentage representations for critical metrics. This normalization allows us to assess gentrification impacts relative to the total population per tract, ensuring comparability across diverse geographic regions. Changes in key socioeconomic indicators over time (2015-2020) were computed to capture the dynamics of

gentrification. These include changes in poverty levels, educational attainment, and racial demographics.

2.Amenity Accessibility:

We conducted spatial joins between census tracts and points of interest to count the number of amenities within each tract. This process included grocery stores, housing permits (indicative of development activity), restaurants, and educational facilities, providing a lens into the changing infrastructural landscape which often accompanies or signals gentrification.

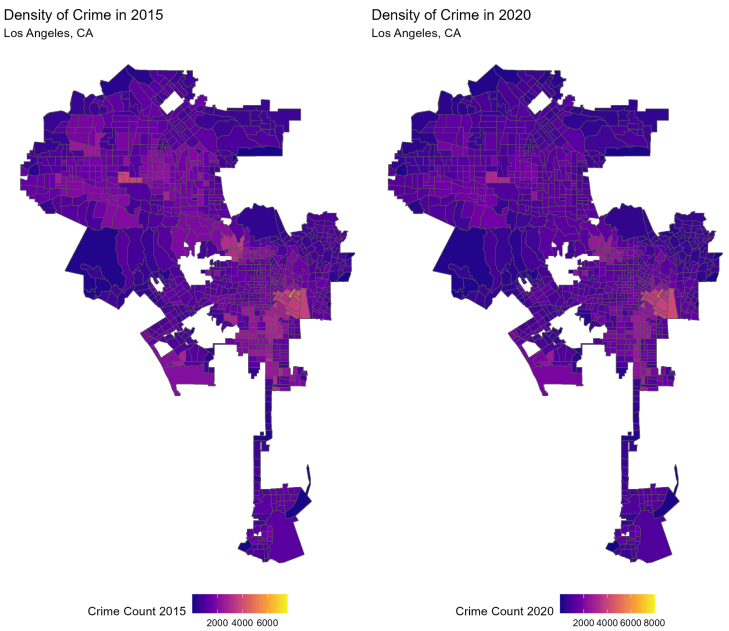


3.Transit-Oriented Development (TOD) Area Identification:

Utilizing "unioned buffer-intersection" techniques, we defined TOD areas by creating a buffer around transit stations and then identifying intersecting census tracts. These tracts were labeled as TOD areas, serving as a proxy for increased accessibility and potential desirability, factors known to influence gentrification.

4. Crime Data Integration:

To understand the relationship between safety and gentrification, we compiled crime statistics for 2015 and 2020, calculating changes over time to discern any correlations between gentrification and crime rates. The following maps illustrate the change in crime density across Los Angeles from 2015 to 2020, it can be observed a decrease in crime density in the upper town area



within the tracts with relatively high crime numbers per tract.

This rich dataset allows for a nuanced analysis of how various factors contribute to or are affected by gentrification, providing city planners with actionable insights into urban development processes.

Predict with Logistic Regression

1. Logistic Regression

As we measured gentrification with binary classification, we chose logistic regression to predict gentrification in Los Angeles because it effectively models the probability of each census tract becoming gentrified based on socio-economic and demographic predictors. We finally involved a list of demographic and socio-economic predictors, include:

- **gentrification:** An index using 0 and 1 to indicate whether a tract has gentrified. 0 for no and 1 for yes.
- **incomeChange:** income change from 2015 to 2020
- **changeinpovwerty:** Change in percentage of poverty from 2015 to 2020
- **ForMig_Change:** Change in percentage of migration from 2015 to 2020
- **changeinbachelor:** Change in percentage of people with at least a bachelor degree from 2015 to 2020
- **housingprice20:** Median rent price per tract in 2020
- **changeinhouseprice:** Change in median house price per tract from 2015 to 2020
- **changein2544:** Change in population aged from 25 to 44 per tract from 2015 to 2020
- **rent20:** Median rent price per tract in 2020
- **rent:**Median rent price per tract in 2015
- **changeinwhite:** Change in percentage of people with at least a bachelor degree
- **pctBachelors20:** Percentage of people with at least a bachelor's degree in 2020
- **newhousingunit:** new housing units from 2015 to 2020
- **Race:** the racial context
- **crimeChange:** Crime change number between 2015 and 2020.
- **Density_Change:** Change in population density between 2015 and 2020

While TOD and amenity features don’t contribute well to the model accuracy so we dropped them in the end. The detailed model table can be found in the appendix.

The final model has an accuracy of over 0.85 and a sensitivity of 0.87, which indicates this model can correctly predict 85% of census tracts are gentrified or not, with an emphasis on correctly predicting

the tracts that got gentrified. With the capability of identifying the areas with gentrification risk, it becomes possible to reduce displacement with in-time affordable housing incentives and better preserve the local culture and local resources. The model has a McFadden score of 0.48, suggesting a good prediction performance. The AUC curve for our model is .83, proposing that we have a strong model with the feature engineered variables.

2.Cross Validation and Generalizability

We further examined the generalizability of the model under the racial context. There are more dominantly minority population tracts that got gentrified between 2015 and 2020 than the tracts with dominantly white populations. Our model has higher accuracy when predicting the dominantly white census tracts, especially when identifying the tracts that are not at the risk of gentrification, while this model performs better in correctly predicting the minority tracts that are at gentrification risk. However, this model generally over-predicts for both dominantly white and minority census tracts.

3.Testing Model with Chicago Data

The validation of our predictive model using Chicago's data, spanning the same years as the Los Angeles dataset, serves as a strategic choice to test the model's applicability and robustness across different urban settings. Choosing Chicago for validating our gentrification model developed from Los Angeles data is predicated on several factors:

1. Similar Urban Dynamics: Both Chicago and Los Angeles are major metropolitan areas with diverse populations and significant economic disparities across different neighborhoods.
2. Prevalence of Gentrification: Like Los Angeles, Chicago has experienced noticeable gentrification, particularly in neighborhoods close to the city center and along key transit routes.
3. Comprehensive Data Availability: Chicago, similar to Los Angeles, has extensive data resources on demographics, urban infrastructure, and socio-economic indicators, which are crucial for a fair comparison and reliable validation of the predictive model.

This approach helps ascertain the model's generalizability and accuracy in predicting gentrification beyond the initial city, ensuring that the model can potentially be adapted for various urban environments with similar underlying patterns of change. The result of this step of validation shows that the model

Recommendation for implementation

1. Projecting Five-Year Scenarios

The model is calibrated to predict changes based on a five-year historical data pattern. Therefore, it is best suited for forecasting similar five-year future intervals. To leverage its predictive power effectively, we recommend utilizing it to project outcomes from 2020 to 2025, and similarly for subsequent five-year periods. This approach aligns the model's strengths with its intended application, ensuring relevance and accuracy in its predictions.

2. Regular Data Updates

For the model to remain effective, it is crucial to incorporate the latest data regularly. As urban demographics and economic conditions evolve, updating the dataset for the next five-year span (e.g., 2025 - 2030) as soon as new data becomes available will enable the model to capture emerging trends and shifts in urban development. This continuous updating process not only enhances the model's accuracy but also maintains its relevance in dynamic urban planning contexts.

3. Model Customization for Highly Gentrified Cities

Cities that have undergone significant gentrification, such as New York, present unique challenges that may not be fully addressed by a general model. In these cases, it is advisable to develop specialized models that consider specific local factors and the saturation of gentrification effects. Such tailored models should focus on more granular aspects of change, such as shifts in micro-neighborhood demographics or the impact of policy changes, to provide useful insights for urban planners and policymakers.

4. Incorporating Localized Factors

Consider enhancing the model by integrating more localized factors that influence gentrification, such as zoning laws, public transportation developments, and economic incentives. This addition can improve the model's ability to forecast gentrification impacts more accurately within specific contexts.

Conclusion

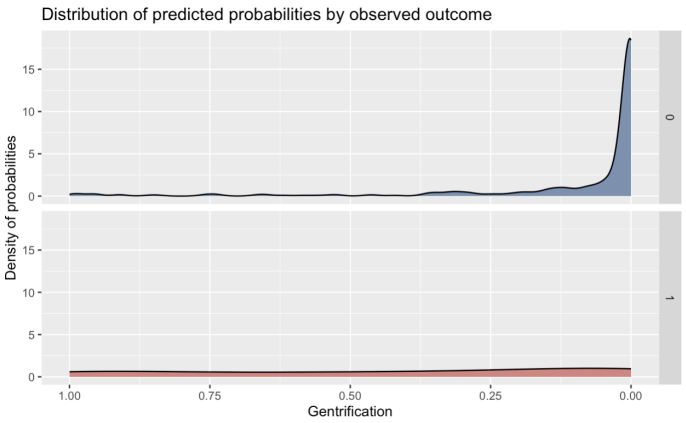
Introduce the predictive model as a state-of-the-art tool designed to identify and forecast gentrification trends within urban areas. Highlight that this model leverages both spatial and non-spatial data from the American Community Survey, covering demographic shifts and socio-economic changes from 2015 to 2020.I would recommend the city to invest in the development of this model to improve the distribution of socioeconomic resources.

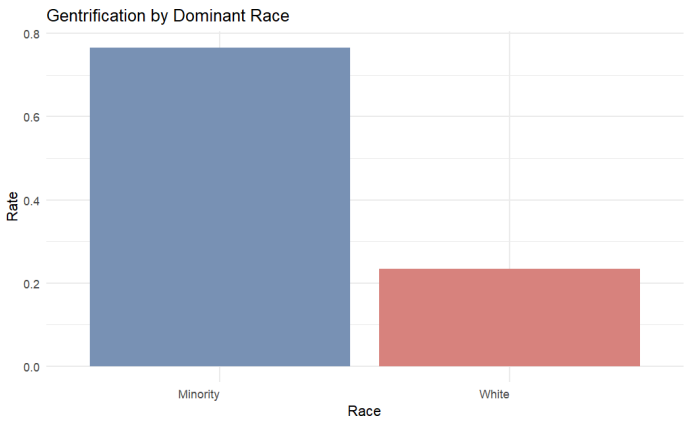
References

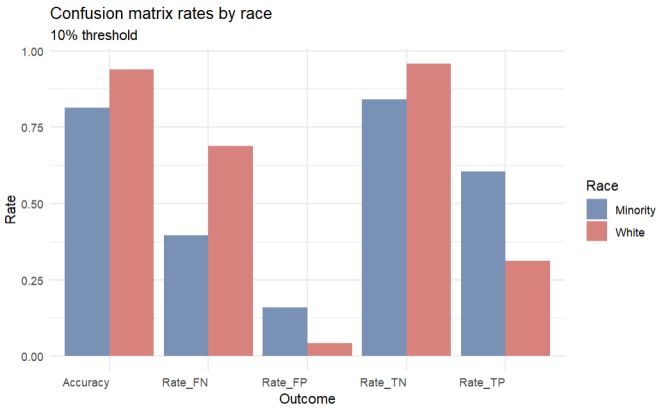
DeVylder, J., Fedina, L., & Jun, H.-J. (2019). The Neighborhood Change and Gentrification Sc
ale: Factor analysis of a novel self-report measure. Social Work Research, 43(4), 279–284.
https://doi.org/10.1093/swr/svz015
Hwang, J. (2016). The social construction of a gentrifying neighborhood: Reifying and redefining identity and boundaries in inequality. Urban Affairs Review, 52(1), 98-128.
Hwang, J. (2015). Gentrification in changing cities: Immigration, new diversity, and racial inequality in neighborhood renewal. The Annals of the American Academy of Political and Social Science, 660(1), 319-340.

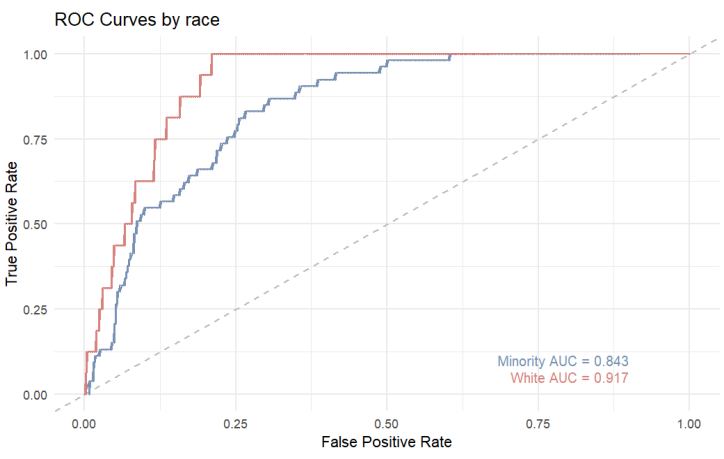
Hwang, J., & Sampson, R. J. (2014). Divergent pathways of gentrification: Racial inequality and the social order of renewal in Chicago neighborhoods. American sociological review, 79(4), 726-751.
Richardson, J., Mitchell, B., & Edlebi, J. (2020). Gentrification and disinvestment 2020.

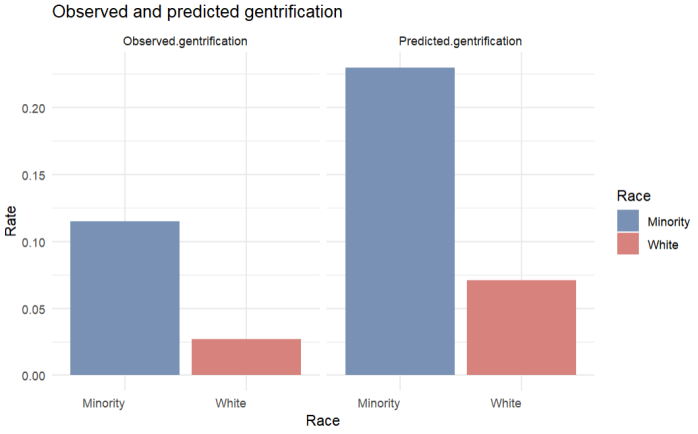
Appendix











6.1 Modeling building

	Estimate	Std. Error	z value
(Intercept)	3.3638534	4.4438213	0.7569731
changeinpoverty	11.5353439	15.6267896	0.7381775
changeinbachelor	-28.5774509	230.2222616	-0.1241298
changeinwhite	0.0412974	6.1224493	0.0067452
pctBachelors20	-190.9230455	357.3621565	-0.5342565
changein2544	-11.7804180	38.6451691	-0.3048355
houseprice20	-0.0000337	0.0000186	-1.8091612
rent	-0.0029736	0.0051635	-0.5758938
rent20	0.0011469	0.0027592	0.4156756
changeinhouseprice	0.0000674	0.0000325	2.0714647
newhousingunit	-0.0053349	0.0092816	-0.5747883
incomeChange	0.0001137	0.0001203	0.9453091
crimeChange	-0.0126408	0.0232188	-0.5444215
ForMig_Change	29.5252539	76.2806294	0.3870610
Density_Change	-755.1584616	8317.1253333	-0.0907956
raceWhite	0.9545769	1.7935846	0.5322174

7.4.3 Confusion matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 189   5
##          1  37  13
##
##                Accuracy : 0.8279
##                  95% CI : (0.7745, 0.873)
##     No Information Rate : 0.9262
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3072
##
##  Mcnemar's Test P-Value : 0.000001724
##
##             Sensitivity : 0.8363
##             Specificity : 0.7222
##          Pos Pred Value : 0.9742
##          Neg Pred Value : 0.2600
##              Prevalence : 0.9262
##          Detection Rate : 0.7746
##    Detection Prevalence : 0.7951
##       Balanced Accuracy : 0.7793
##
##        'Positive' Class : 0
##
```

```
##
## Call:
## glm(formula = gentrification ~ ., family = binomial(link = "logit"),
##     data = gentrifydataTrain %>% dplyr::select(gentrification,
##         changeinpoverty, changeinbachelor, changeinwhite, pctBachelors20,
##         changein2544, houseprice20, rent, rent20, changeinhouseprice,
##         newhousingunit, incomeChange, crimeChange, ForMig_Change,
##         Density_Change, race))
##
## Coefficients:
##                         Estimate    Std. Error z value Pr(>|z|)
## (Intercept)           3.36385339    4.44382128   0.757   0.4491
## changeinpoverty      11.53534393   15.62678957   0.738   0.4604
## changeinbachelor    -28.57745090  230.22226158  -0.124   0.9012
## changeinwhite         0.04129735    6.12244932   0.007   0.9946
## pctBachelors20     -190.92304553  357.36215654  -0.534   0.5932
## changein2544        -11.78041802   38.64516911  -0.305   0.7605
## houseprice20         -0.00003367    0.00001861  -1.809   0.0704 .
## rent                 -0.00297365    0.00516353  -0.576   0.5647
## rent20                0.00114693    0.00275920   0.416   0.6776
## changeinhouseprice    0.00006742    0.00003255   2.071   0.0383 *
## newhousingunit       -0.00533493    0.00928155  -0.575   0.5654
## incomeChange          0.00011369    0.00012026   0.945   0.3445
## crimeChange          -0.01264079    0.02321875  -0.544   0.5862
## ForMig_Change        29.52525395   76.28062941   0.387   0.6987
## Density_Change     -755.15846164 8317.12533330  -0.091   0.9277
## raceWhite             0.95457694    1.79358457   0.532   0.5946
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 46.13  on 72  degrees of freedom
## Residual deviance: 23.92  on 57  degrees of freedom
##   (43 observations deleted due to missingness)
## AIC: 55.92
##
## Number of Fisher Scoring iterations: 9
```

The model has a McFadden score of 0.48, suggesting a good prediction performance.

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -11.9597969 -23.0649552  22.2103165   0.4814732   0.2623242   0.5600098
```