Predicting Gentrification via Socioeconomic Changes:

Evidence from New York City 2012-2019

FNU Sonam (ss15624), Max Magid (mmm9940), Alec Bardey (ab9732), Suraj Sunil (ss14449)

Abstract

The definition of gentrification is inherently ambiguous, where some understand it as urban renewal, others are directly impacted by being priced out of neighborhoods. Past literature has covered the use of neural networks (Ilic et al., 2019), clustering algorithms (Liu et al., 2019), and supervised learning methods (Thackway et al., 2021). Few of these studies use widely available and easily scalable data and few evaluate New York City. We use a k-means clustering approach on changes in American Community Survey data from 2012-2019 to create gentrification labels. We then use these labels in addition to indicators for spatially autocorrelated hotspots to predict gentrifying tracts on absolute values of ACS data. Our unsupervised analysis revealed easily explainable regions that fit with anecdotal evidence of gentrification. By pairing our most precise method (random forest) with our method with the best recall (PCA regression), we can predict gentrifying regions with 40% accuracy, lower-likelihood regions with 20% accuracy, and have a resulting total recall of 40%. These results significantly outperform the baseline of 10% precision.

Problem Statement

Identify, characterize and locate neighborhoods in New York City which have undergone recent gentrification, specifically disaggregating the different types of change revealed by the data (from 2012 to 2019). Explore which neighborhoods are likely to be next in line. Present and make available data, code, and novel interactive visualizations as a comprehensive tool for supporting policy- and decision-making in the city.

Introduction

The term ‘gentrification’ can take on many meanings and interpretations. Many understand it to represent a natural process of urban renewal characterized by rent increases, neighborhood beautification, and a wealthier population. At its core, however, gentrification is the indirect displacement of lower-income residents from a community via the migration of wealthier citizens seeking savings on housing costs. Landlords and business owners capitalize on the influx of wealth by raising rents and prices, which often leads to the ‘pricing-out’ and displacement of local residents from the neighborhood. While this can lead to prettier streets and nicer schools, few of the original residents reap the benefits.

Although de jure racial segregation was outlawed 60 years ago, de facto segregation persists in cities through sanctioned redlining and modern gentrification. Gentrification exists globally but manifests uniquely in America, where these practices were historically codified. Many view gentrification as a modern form of colonialism with negative social, economic, and health outcomes for disadvantaged communities. Socially, gentrification has been found to perpetuate racial and ethnic inequality (Hwang, 2015). Economically, the mobility patterns of economically disadvantaged residents are unequally distributed by race (Hwang & Ding, 2020). In other words, displaced residents have diminishing options and opportunities for new residencies and jobs. Finally, gentrification has been found to lead to statistically worse health outcomes for Black and Latino citizens as these communities are pushed out of areas with access to better facilities (Gibbons & Barton, 2016).

In America's densest metropolitan area, New York City, gentrification is especially prevalent and volatile. Specific studies in New York have corroborated the general trends outlined above. An analysis of census and retail data from 1970-2010 found that gentrification occurs at higher rates in areas with higher proportions of Black and Latino residents (Sutton, 2020). The study finds that even middle-class minorities face difficulties in remaining in actively gentrifying neighborhoods. Another study evaluated the causal relationships of gentrification in NYC from 2009-2016 and found that a 1% increase in gentrifiers was associated with a 2.7% increase in property values, yet the same relationship was not observed in reverse (Zapatka & Beck, 2021).

Identifying and potentially predicting gentrifying tracts can help develop equitable housing solutions. However, the inherent definitional and practical ambiguity of gentrification can make identification particularly difficult. Can we use the causal relationship between gentrifying populations and rising housing properties outlined by Zapatka & Beck to guide gentrification recognition using publicly available data? This project aims to use unsupervised machine learning methods to identify gentrifying tracts in New York City. Using these classifications, we then aim to build a model to predict what tracts will gentrify and what features most heavily contribute to gentrification in New York City from 2012-2019.

Background

Gentrification has been studied from a multitude of technical and sociological perspectives. Above we outlined some sociological implications and impacts of gentrification in urban areas. Here we will focus on the technical research in identification and prediction.

One recent method explored in the literature involves the use of visual components as gentrification predictors. Ilic et al. (2019) use a convolutional neural network to detect ‘gentrification-like’ physical and visual changes to neighborhoods. By analyzing individual properties as opposed to general neighborhood dynamics, they achieve high test accuracy for gentrification detection in Ottawa. Lin et al. (2021) compare a multitude of deep learning methods. They describe the moderate success of remote sensing via coarse satellite images in categorizing properties. They further find that Call Details Records have been used as poorer areas in cities tend to have fewer class but higher Short Message Service traffic. Finally, the researchers outline that there lacks scalability with a great deal of these deep learning approaches and indicate the use of ACS data as a potentially scalable and low cost identification mechanism.

A slew of other papers use more traditional machine learning approaches in gentrification analyses. Jain et al. (2021) explore the use of unstructured and structured AirBnB data to identify ongoing gentrifying trends in New York City, Los Angeles, and London. They find that unstructured data like user-generated reviews can be used to predict neighborhood demographics and housing prices. This approach is expensive, as well as sensitive to changes in the company’s practice and data accessibility policies. Liu et al. (2019) compare a threshold method with a k-means clustering approach to identify gentrifying tracts in Auckland. The threshold method was deemed slightly more accurate. While we ultimately use a clustering approach, we use more features with a Principal Component Analysis (PCA) for clustering identification. Furthermore, we wanted to avoid subjective and confirmation bias so we neglected to take a universal threshold approach for NYC.

Clustering methods to identify gentrifying tracts in urban areas have observed success. Corrigan et al. (2021) evaluate spatial autocorrelation of gentrification in Atlanta from 2000-2016. They use PCA to develop a single gentrification score, which they then use to measure spatial hotspots. We have similarly elected to use PCA to get a gentrification score, but we then use dummy values for significant spatially autocorrelated hotspots to guide our supervised learning approach. Ferdowsi et al. (n.d.) cluster Chicago census tracts on 5 target variables. They do not compute accuracy, but rather evaluate the proportion of Chicago community areas that belong to each cluster. Knorr (2019) uses k-means to identify 4 clusters in Nashville, TN from 6 change variables. Similar to Corrigan et al., Knorr follows this with spatial statistics analysis and concludes that there is potential to predict gentrification with these methods.

There is also evidence in the literature regarding the success of supervised learning and prediction methods. Reades et al. (2019) use PCA on 4 features on UK housing information from 2001-2011 to guide a Random Forest approach to help predict what tracts will gentrify by 2021. We intend to combine a version of this approach with unsupervised clustering to objectively identify gentrifying tracts. Thackway et al. (2021) also use a Random Forest to model gentrification in Sydney from 2011-2016. They use this model to predict gentrification through 2021. This paper uses absolute values for specific years, as opposed to changes over time. Alejadrio & Palafox (2019) use a Random Forest approach to evaluate gentrification in Mexico

City from Census data from 2000-2016. The core feature variable used in this analysis was the percentage of people in a given neighborhood living in a different city in years prior.

Yee & Dennett (2021) use an approach most similar to our outlined methodology. The researchers use census data in London from 2001-2021 to predict gentrification. They include voter information, tax information, and land use information to guide a PCA to create a composite index for gentrification. They follow this with a threshold metric to create 3 classifications of gentrification and use these classification to guide a Random Forest prediction model

Gentrification is so spatially, historically, and culturally dependent that different contexts and datasets can produce different results. This underscores the importance of using scalable and widely available data when evaluating gentrification. For this reason, this paper uses census demographic information to guide gentrification identification. There are no papers that have used a combined clustering, PCA, and Random Forest approach to identify and predict gentrification in New York City. We also further the methodology described above as we use change over time as the mechanism for clustering gentrifying regions, then use absolute values as potential predictors. This represents a clear difference in the methodology employed by Thackway et al. (2021) and Reades et al. (2019). The approach is somewhat similar to the method used by Yee and Dennett (2021) but our paper fills a gap by using more scalable data (American Community Survey), an understudied region (NYC), and using spatial autocorrelation measures as features in prediction

Data

The data used in this project is an abstract from the United States Census Bureau tract of geodatabases from 2012 to 2019. The data selected for evaluating gentrification from the census tract has geolocation of housing units in New York city and corresponding data on income level, rent prices, ownership and vacancy rates, ethnicity of households residing, poverty and education level. The dataset has nearly 2100 records of household units in each year.

The detailed description of each label in the dataset can be seen in the following table:

Label	Description
GEOID	Locations ID
year	Year
m_income	Median Income of households
m_rent	Median Rent
m_own	Median property price of location ID
perc_white	Percent White

perc_poverty	Percent Poverty
perc_o25_ed	Percent over 25 with education
owner_occ	Percent housing occupied by owners
rent_occ	Percent housing occupied by renters
rent_vac	Percent vacant Rental property
temp_vac	Percent of temporary housing (seasonal)

Methodology

Diverse methods, ML and spatial analysis, were chained into a multi-level workflow to unpack the diversity and trajectories of gentrification across neighborhoods in New York City. Firstly, statistical geographies in New York City were classified according to whether they were ascending, stable or declining in terms of house price, levels of highly educated, high socio-economic status residents or median income. These classifications were then used to predict future areas of gentrification in the city using machine learning algorithms.

This research conceptualizes gentrification as a change in the economic, racial, and age composition of an area such that the community experiences not only an increase in economic privilege, but may also experience an increase in other markers of privilege, including Whiteness, youth and occupancy of spaces. A variety of neighborhood variables are compiled into a single gentrification score representing the overall degree of gentrification in each geography between two time points. Principal component analysis (PCA) is a common data reduction technique used to create indices that summarize the total area-level variance explained by a large number of indicators. While used to summarize economic disadvantage in several studies (Genberg et al., 2011; Walker et al., 2020)., few studies have specifically used PCA to develop indices of gentrification.

Census tract economic and other demographic data, including race/ethnicity and age,were acquired from the U.S. Census Bureau Decennial Census. Ten variables were included as part of the gentrification score: 1) Median household income 2) Median rent 3) Median value of owner-occupied property 4) Percentage of people who are non-Hispanic white 5) Percentage of households living below poverty line 6) Percentage of population over 25 with some college or more 7) Amount of housing occupied by owner 8) Amount of housing occupied by renters 9) Amount of vacant rental property 10) Amount of temporary housing (airbnb, seasonal workers, etc).

A gentrification score was calculated by summing the standardized percent change between time 1 (t1) and time 2 (t2) for all indicators. These t1 and t2 (2012-2015,2014-2017,2016-2019) are chosen based on previous exploratory analysis and research. A positive value for any indicator identified a change greater than the metro-wide trend and stronger relative gentrification. After taking the sum of all indicators, the composite gentrification scores were standardized again so that a score of 0 represents average gentrification for the entire study area.

To establish interpretable gentrification scores when indicator values at t1 and t2 both equaled zero, 0 was imputed to reflect no change. When indicator values equal zero at t1 but were non-zero at t2, a small number was imputed (1 added to each value) so that the positive change in the given gentrification component would be counted. If an indicator was missing in either t1 or t2, the percent change was NA. Composite gentrification scores were calculated by imputing the missing data. Gentrification scores using the 10 indicators were first calculated for 2012–2015. Next, to more closely examine population and economic trends in early versus later stages of gentrification, scores were calculated for two more periods, 2014–2017 and 2016–2019.

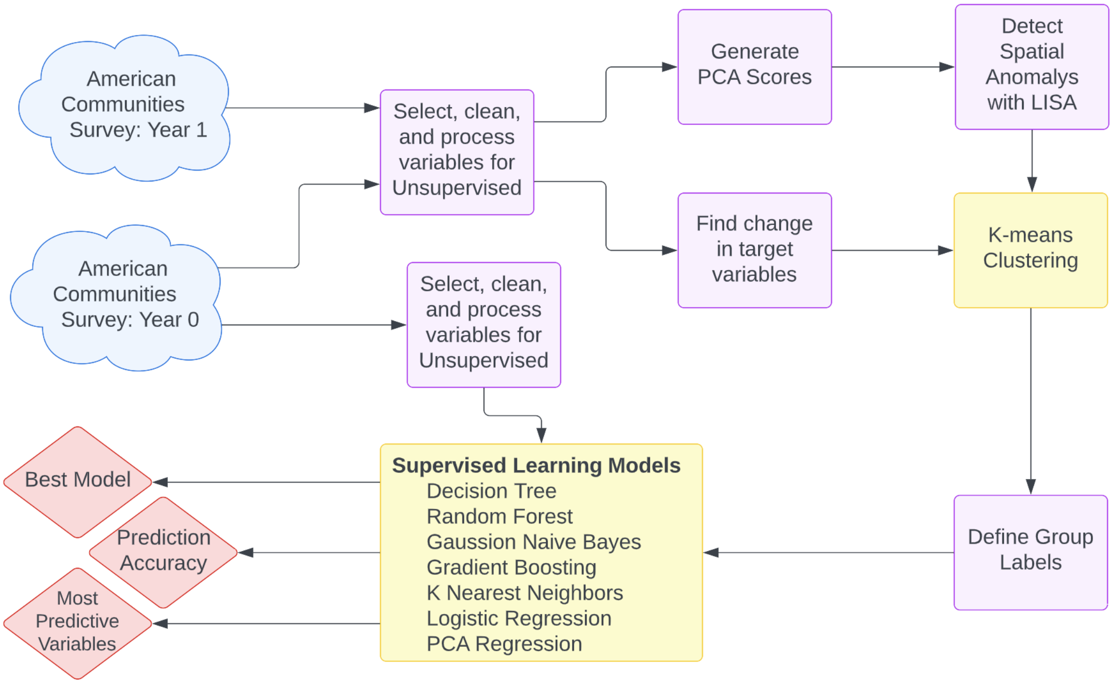
American Communities Survey: Year 1
Select, clean, and process variables for Unsupervised
Generate PCA Scores
Detect Spatial Anomalys with LISA
American Communities Survey: Year 0
Find change in target variables
K-means Clustering
Select, clean, and process variables for Unsupervised
Supervised Learning Models
Decision Tree
Random Forest
Gaussion Naive Bayes
Gradient Boosting
K Nearest Neighbors
Logistic Regression
PCA Regression
Define Group Labels
Best Model
Prediction Accuracy
Most Predictive Variables

Figure 1: Methodology flow chart

We then took the changing score between t1 and t2 to calculate local Moran's I statistics to find spatial autocorrelation. Keeping only statistically significant measures we added information on whether the tract was negatively changing and surrounded by other negatively changing areas (low-low), was a positively changing tract surrounded by other positively changing tracts (high-high), or if it was an outlier (high-low, low-high). We created dummy variables and attached this information to our unsupervised data frame as four new columns.

A k-means algorithm was used to isolate clusters of neighborhoods with similar characteristics. We then analyzed the clusters using average values from the group, as in the radial plot below. Through this we identified fitting names for the three clusters: Gentrifying, Stationary and Declining.

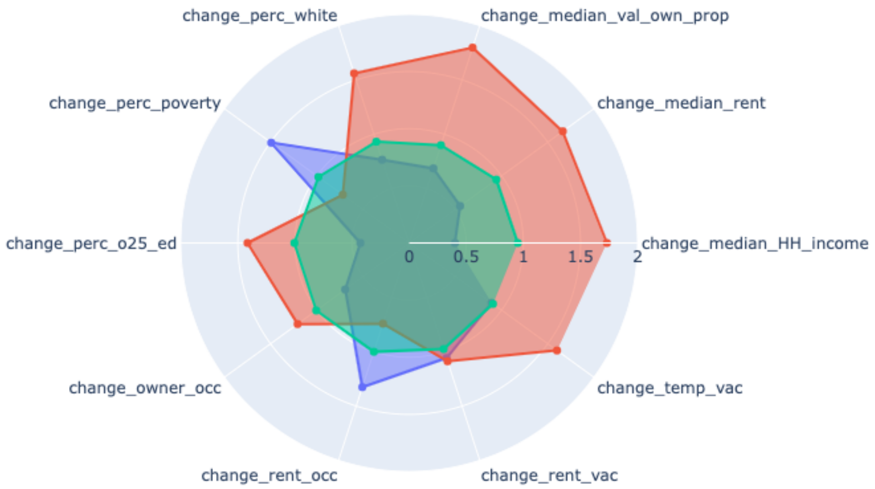

Figure 2: Wheel chart representing cluster characteristics

We trained a suite of Machine Learning models on the observed trends and spatial patterns of neighborhood ascent and gentrification that unfolded between 2012-2019, with the aim of predicting areas that will gentrify in the near future and their corresponding typologies. To further optimize the model, We added additional demographics such as median age, sex, and races other than white. We also included spatial autocorrelation through local Moran's I based on the PCA score for year 0 in our analysis. Furthermore we experimented with adding latitude and longitude of the center points of each tract, or including the borough each tract fell into. We then trained a variety of machine learning models on the data to see which would perform best. We tested random forest, gaussian naive bayes, a decision tree, gradient boosting, k nearest neighbor, logistic regression, and pca regression.

Results

We found that the PCA regression model had the best recall, but the random forest model had the best precision. Both performed 3-4 times better than random chance.

Furthermore, by tuning the random forest regression we were able to find the best possible random forest result based on our data and the most important features. We found that a random forest model using boroughs instead of latitude/longitude with 98 estimators and a max depth of 15 lead to a model with precision nearly four times as high as the baseline and recall higher than any model besides pca regression.

Figure 3 below represents the significance of each variable in the Random Forest. As we can clearly see, the percentage of people having age above 25 and are educated is the most significant feature. This is followed by median rent and median household income and many more.

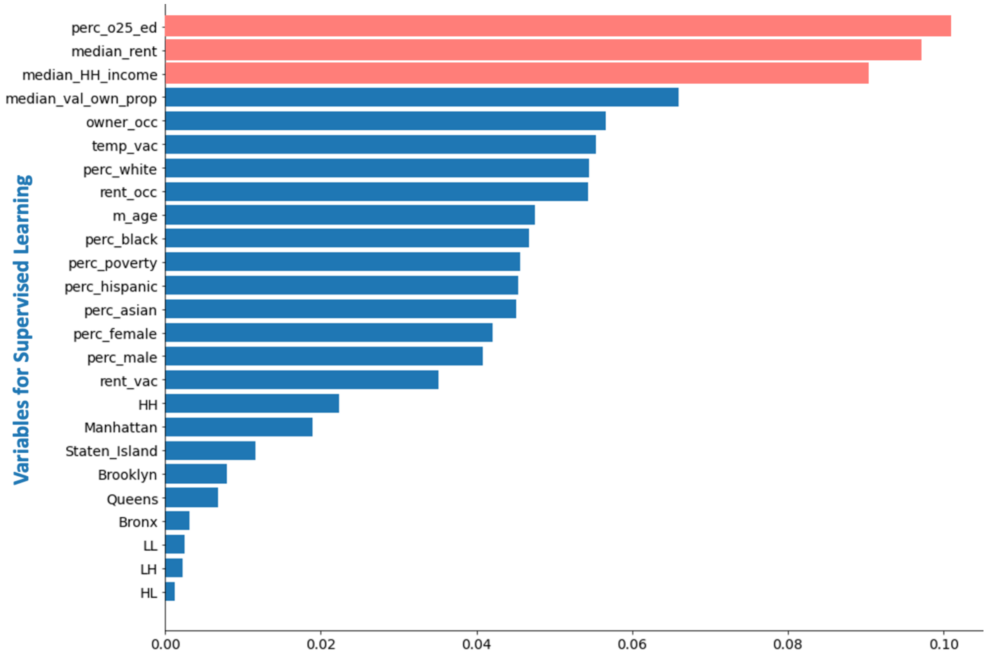

Figure 3: Feature importance for Random Forest

Figure 4 shows the spatial statistics based on location and the composite feature score found using principal component analysis (HH, LL, HL, and LH) were not particularly helpful in predicting gentrifying tracts. However, using a similar method on the unsupervised dataset strongly impacted the clusters formed. Below is the visualization for the LISA statistics around the difference in pca value between year 0 and year 1. Below that is the heavily correlated plots generated from our k-means clustering.

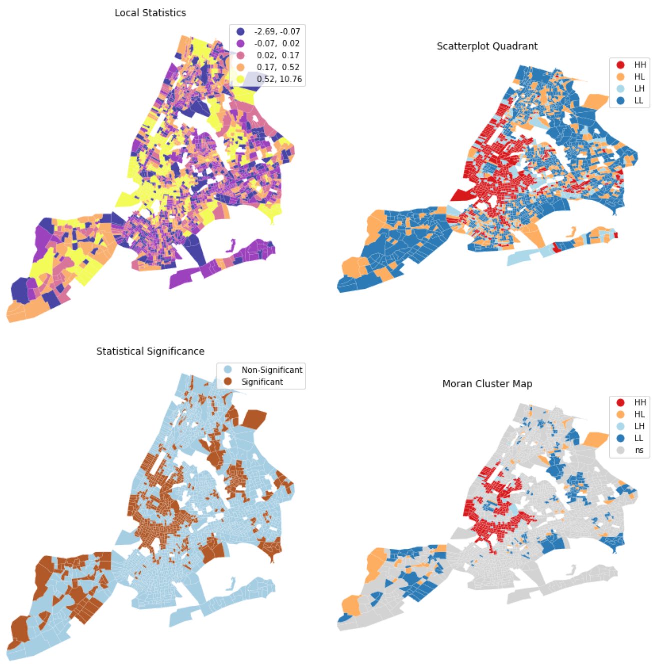

Figure 4: Spatial statistics by location using PCA

The top map of Figure 5 shows the areas which have gentrified from 2012 to 2015. Red color represents the gentrifying area and is more concentrated around Manhattan. This shows that Manhattan has transformed a lot in the past three years from 2012 to 2015 with exponential growth in industry and people moving in the area for permanent settlement. Green represents the declining areas where gentrification is declining, as we can clearly see the areas in Staten Island which were gentrifying earlier and are now declining. Gray color is for stationary areas where gentrification has had no impact so far but can be under gentrification in the coming future.

In the 2019 plot of Figure 5, we can see the clear change in areas which are gentrifying. The 2019 plot shows that Manhattan reached its peak of gentrification in previous years until 2019 and now stationary. The urban sprawl is moved to outer areas like Brooklyn, and we can clearly see that downtown areas of Brooklyn are under maximum gentrification in 2019. The area of Staten Island has turned gray from green, which shows that declining has reached another extreme and has become stationary. Please refer to - *gentrification_comparison.html* for interactive map.

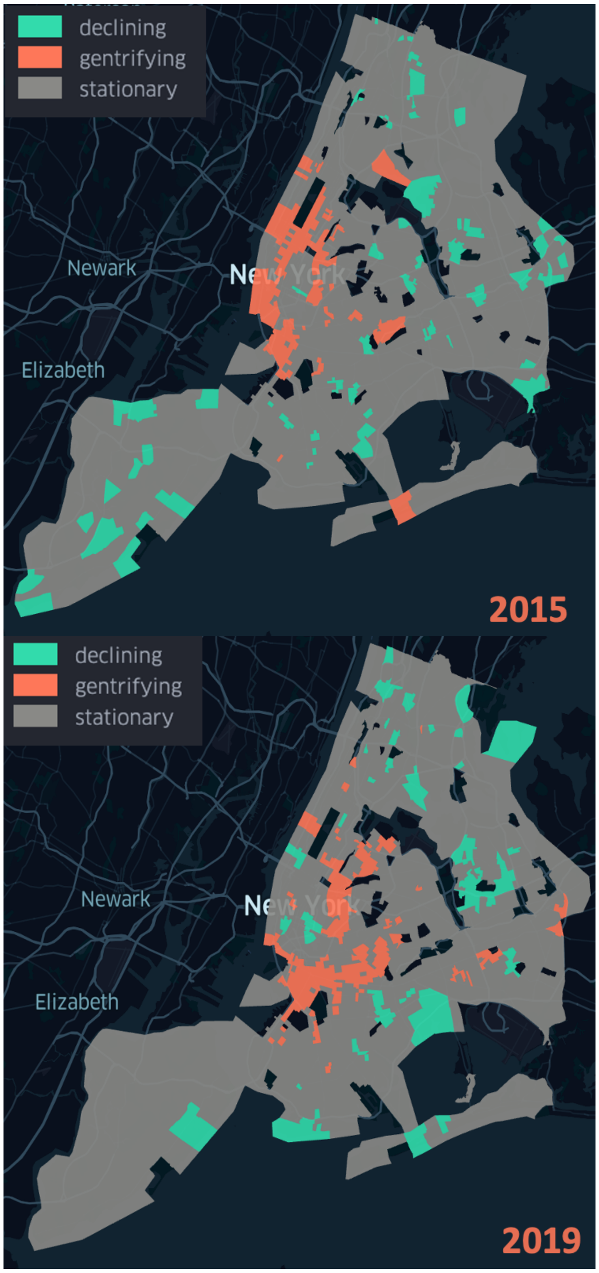

Figure 5: Census tract classification results for 2015 and 2019 (using KeplerGL)

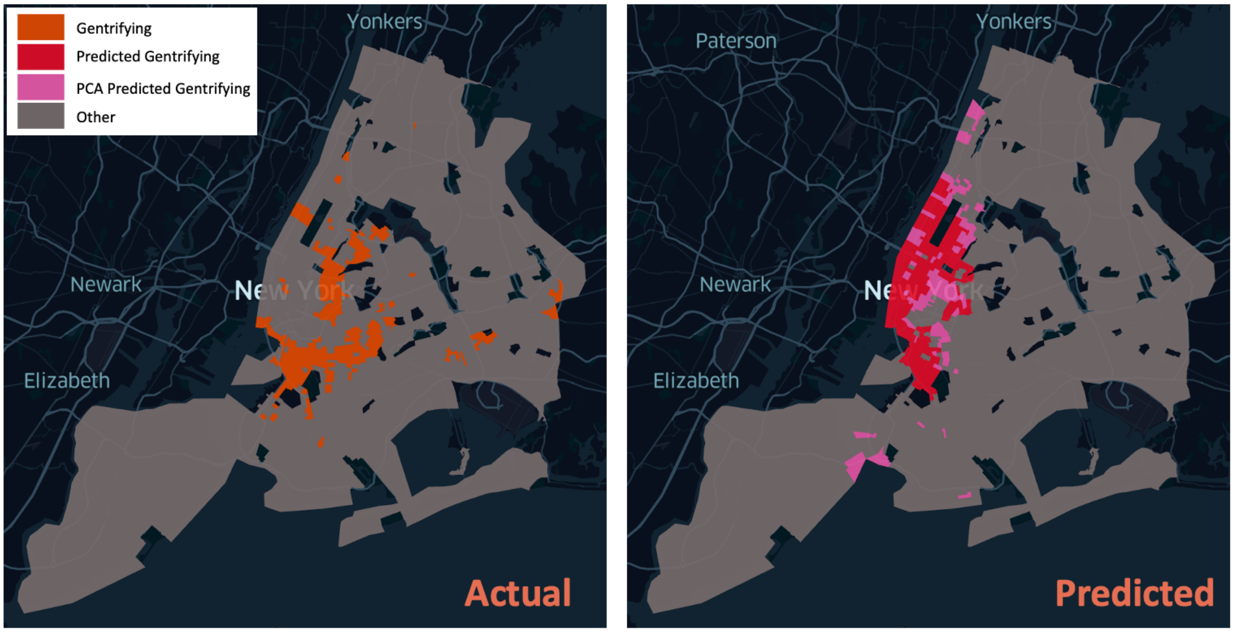

Figure 6: Actual vs. predicted gentrification clusters from Random Forest (using KeplerGL)

Our predicted values largely missed the actual shift away from Manhattan and into Brooklyn. The red areas have a predictive precision of 38% while the pink areas have a predictive precision of 20%. Together the areas cover nearly 40% of actually gentrifying areas. Please refer to - *Supervised_Learning_output_Analysis.html* for interactive map.

Discussion

The research presents a methodology to identify gentrification areas in New York city from 2012 and 2019. We identified major parameter change typologies via k-means clustering of important census tract variables. We classified the locations according to stable, gentrifying and declining in terms of all the features labels present in the data. The distinct markers of gentrification were mostly signaled by the increase in educational attainment of population above 25, home value and income group. These values strongly suggest residential displacement within gentrifying areas within the period. The areas experiencing declining gentrification might be potentially the inverse of the signaled gentrification implying increased concentration of poverty rate in the locations.

Gentrification can be predicted by studying the trajectorial change in the characteristics of neighborhoods. Yet, upgrades and downgrades of a neighborhood can be distinguished and can be subject to change according to the characters causing it. Different cities have different characteristics and processes of gentrification, therefore developing a machine learning framework that best captures the cause are relative to change according to the methods followed to predict gentrification. The research demonstrates the performance of different supervised machine learning methods to identify areas of future gentrification in the city. The areas seen by increase in home values, income and education levels are predicted to have the highest probability of gentrifying. The study finds that these factors will have the greatest impact on how gentrification will manifest in future years in New York city.

Conclusions

The first stage of the research focuses on identifying change in patterns of characteristics in the city through the years using an unsupervised machine learning framework defining the regions in the city by trends of declining, stationary and gentrifying. Neighborhoods in the city were scored according to the degree of gentrification using Principal Component Analysis. The process was followed for selected years between 2012 to 2019 and a composite gentrification score was computed to identify areas of significant difference in two different time frames using Moran's I statistics. K-means algorithm was further used to classify the neighborhoods as gentrifying, stationary or declining. Further stages of the research involved identifying the best practices to predict gentrification using supervised learning machine learning approaches.

The results showed that results obtained from fine tuning random forest and PCA regression model provided satisfactory results in predicting gentrification, both demonstrating best results on precision and recall respectively. Executing both the models results will provide the potential areas which will experience gentrification in future years. Though this study provides a machine learning framework to identify gentrification in cities, further research requires studying more characteristics which might have a greater impact on gentrification. Future studies may also benefit from additional features in the supervised learning steps like land use information from PLUTO data, transportation information, accessibility, and available amenities.

Team Contribution

- **FNU Sonam (ss15624):** Data Pre-Processing, PCA analysis and scores, k-means clustering, KeplerGL interactive visualizations, Methodology for unsupervised learning
- **Max Magid (mmm9940):** Methodology- LISA statistics, PCA regression, fine-tuning random forest model
- **Alec Bardey (ab9732):** Literature review, introduction + background, creating training and testing datasets, random forest model
- **Suraj Sunil (ss14449):** Conclusions, data section, Comparing random forest, Gaussian Naive Bayes, decision tree, k nearest neighbor, and boosting methods

References

1. 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.
2. Hwang, J., & Ding, L. (2020). Unequal displacement: gentrification, racial stratification, and residential destinations in Philadelphia. *American Journal of Sociology*, *126*(2), 354-406.
3. Gibbons, J., & Barton, M. S. (2016). The association of minority self-rated health with black versus white gentrification. *Journal of urban health*, *93*(6), 909-922.
4. Sutton, S. (2020). Gentrification and the increasing significance of racial transition in New York City 1970–2010. *Urban Affairs Review*, *56*(1), 65-95
5. Zapatka, K., & Beck, B. (2021). Does demand lead supply? Gentrifiers and developers in the sequence of gentrification, New York City 2009–2016. *Urban Studies*, *58*(11), 2348-2368
6. Ilic, L., Sawada, M., & Zarzelli, A. (2019). Deep mapping gentrification in a large Canadian city using deep learning and Google Street View. *PloS one*, *14*(3), e0212814.
7. Lin, L., Di, L., Zhang, C., Guo, L., & Di, Y. (2021). Remote Sensing of Urban Poverty and Gentrification. *Remote Sensing*, *13*(20), 4022.
8. Thackway, W., Ng, M. K. M., Lee, C. L., & Pettit, C. (2021). Building a predictive machine learning model of gentrification in Sydney.
9. Jain, S., Proserpio, D., Quattrone, G., & Quercia, D. (2021). Nowcasting gentrification using Airbnb data. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-21.
10. Liu, C., Deng, Y., Song, W., Wu, Q., & Gong, J. (2019). A comparison of the approaches for gentrification identification. *Cities*, *95*, 102482.
11. Corrigan, A. E., Curriero, F. C., & Linton, S. L. (2021). Characterizing clusters of gentrification in metro Atlanta, 2000 to 2016. *Applied Geography*, *137*, 102597.
12. Royall, E. B. (2016). *Towards an epidemiology of gentrification: Modeling urban change as a probabilistic process using k-means clustering and Markov models* (Doctoral dissertation, Massachusetts Institute of Technology).
13. Ferdowsi, Z., Settimi, R., Raicu, D., & Curran, W. From Individual Tracts to Community Segments: an Unsupervised Learning Approach for the Chicago Community Areas.
14. Alejandro, Y., & Palafox, L. (2019, October). Gentrification prediction using machine learning. In *Mexican International Conference on Artificial Intelligence* (pp. 187-199). Springer, Cham.
15. Knorr, D. C. (2019). *Using Machine Learning to Identify and Predict Gentrification in Nashville, Tennessee* (Doctoral dissertation).
16. Reades, J., De Souza, J., & Hubbard, P. (2019). Understanding urban gentrification through machine learning. *Urban Studies*, *56*(5), 922-942.
17. Yee, J., & Dennett, A. (2021). Stratifying and predicting patterns of neighborhood change and gentrification: An urban analytics approach. *Transactions of the Institute of British Geographers*.
18. Genberg, B. L., Gange, S. J., Go, V. F., Celentano, D. D., Kirk, G. D., Latkin, C. A., & Mehta, S. H. (2011). The effect of neighborhood deprivation and residential relocation on long-term injection cessation among injection drug users (IDUs) in Baltimore, Maryland. *Addiction*, *106*(11), 1966-1974.
19. Walker, A. F., Hu, H., Cuttriss, N., Anez-Zabala, C., Yabut, K., Haller, M. J., & Maahs, D. M. (2020). The neighborhood deprivation index and provider geocoding identify critical catchment areas for diabetes outreach. *The Journal of Clinical Endocrinology & Metabolism*, *105*(9), 3069-3075.