Identifying Gentrification using Machine Learning*

SEHSD Working Paper Number 2023-15

April 20, 2023

Jayne Yoo, U.S. Census Bureau

Introduction

Gentrification has accelerated in cities across the U.S. since the 1990s (HUD, 2018). While the precise definition of gentrification is itself open to debate, discussions of gentrification generally include concerns about the influx of households with higher incomes and more resources into lower-income neighborhoods. Although the accompanying changes may include benefits to some existing residents through improved amenities, increased access to jobs and higher home values (Byrne, 2002; Vigdor et al, 2002; Brummet and Reed, 2019), gentrification may also carry steep costs for residents who cannot afford to stay in the neighborhood due to increased rents, lease refusals, or other reasons. Concerns about these costs have made gentrification a major policy issue in many metropolitan areas, with substantial resources invested in understanding the patterns of gentrification and developing proactive strategies to keep housing affordable.

Since the early 2000's, a large and growing literature has examined the causes and consequences of gentrification, as well as potential response strategies that might reduce the displacement of low-income residents from gentrifying neighborhoods. Despite the abundance of empirical and qualitative studies on gentrification, tools for predicting which residents and neighborhoods will be affected by gentrification remain limited. The importance of early warning systems and predictive empirical models have been discussed as essential tools to facilitate proactive intervention strategies to address gentrification pressures in communities with active real estate development (Chapple & Zuk, 2016).

In this context, this paper explores the potential for machine learning techniques to identify housing units at high risk of gentrification, contributing to the literature on methods for forecasting future gentrification risk. Specifically, the analyses test the performance of multiple alternative machine-

* This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. All errors are those of the authors. Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure disclosure avoidance protection of the confidential source data used to produce this product. Disclosure Review Board (DRB) approval number: CBDRB-FY23-0210.

learning methods, using American Housing Survey (AHS) data for the Washington D.C. Metropolitan Statistical Area (MSA).[1]

The empirical results provide a case study of gentrification in the Washington D.C. metropolitan area, adding to the set of case studies currently available for other cities. According to American Community Survey in 2021, the MSA median housing value of the Washington D.C. metropolitan area is 501,500 dollars whereas the national median price in the U.S is 309,000 dollars. With one of the highest average rental costs compared to other metropolitan areas (Kingsley, 2017)., D.C.'s housing market suffers from housing availability and affordability issues, and it has experienced substantial gentrification in recent years. Population growth and the influx of highly educated workers have driven recent real estate development in Northern Virginia and Maryland that has increased the housing inventory, but housing affordability challenges for low-income residents have also expanded to these areas (Kingsley, 2017). This case study therefore provides evidence about the extent and nature of recent gentrification in a high-cost U.S. city.

Since AHS is a panel data, it allows to identify housing unit turnover and observe detailed socioeconomic characteristics of both the departing and arriving households. The AHS also contains data on housing characteristics and home values that can be used to predict which housing units have higher and lower probabilities of displacement. In addition to the AHS data, the analyses incorporate several novel predictors from a commercial real estate website, showing that these neighborhood characteristics improve the model's accuracy by 23 percent by capturing detailed neighborhood information.

Literature Review

Although many definitions of gentrification exist, U.S. Department of Housing and Urban Development defines gentrification as "a form of neighborhood change that occurs when high-income groups move to low-income areas, potentially altering the cultural and financial landscape of the original neighborhood". Existing literature emphasizes the in-migration of upwardly mobile households into lower socioeconomic areas and the resulting changes to the neighborhood in terms of demographics, land-use, and housing affordability (Lees et al, 2013).

Many empirical studies that examine gentrification discuss socioeconomic changes among residents, displacement, increased investment, and rapid change in neighborhoods. Mayer (1981) and Melchert and Naroff (1987) suggest that property renovation can be a good indicator for gentrification. In terms of neighborhood change, Glaeser et al. (2018, May) uses Yelp data and find that entry of Starbucks into a neighborhood is indicative of housing price growth across the United States and that the number of reviews of Starbucks increases predictive power, suggesting that gentrifying neighborhoods might also attract more reviewers (i.e., neighborhood engagement). O'Sullivan (2005) also shows that decreases in crime or increases in the frequency of travel to city centers can also contribute to increases in gentrification (O'Sullivan, 2005).

[1] A metropolitan or micropolitan statistical area (MSA) is that of a core area containing a substantial population nucleus, together with adjacent communities having a high degree of economic and social integration with that core. Detailed information on MSA can be found here: https://www.census.gov/programs-surveys/metro-micro/about.html

Socioeconomic and demographic changes among residents are widely used as signs of gentrification. Freeman (2005) explores the increase in residents with college degrees and its positive association with housing price growth. Increases in household income over time also indicates gentrification and this measure was used to confirm that gentrification has expanded outwards from the city center (Thackway et al, 2021; Ellen and O'Regan, 2011; McKinnish et al., 2010).

More recently, several gentrification studies have employed machine learning models to capture the complex and dynamic nature of neighborhood change (Reades et al., 2019; Palafox & Ortiz-Monasterio, 2020). The ML methods overcome the limitations of traditional parametric models in modelling social change or behavior. Unlike traditional models, machine learning models can add all features contributing to the model even if they are correlated (Ogutu et al., 2011). Furthermore, interaction between features can be easily addressed using "bagging methods" that aggregate multiple model fits. The non-parametric models make no assumptions about the feature distribution (Ogutu et al., 2011). Tree-based methods can process a large number of observations and features in datasets (Natekin & Knoll, 2013).

Due to the advantages of ML models, researchers have used this technique to identify gentrifying neighborhoods in several case studies (Thackway et al., 2021 for Sydney; Reades et al, 2019 for London; Alejandro & Palafox, 2019 for Mexico). As big data from private sector has become more available, studies also adopted commercial data for ML models. Ilic et al. (2019) uses Google Street View and Jain et al. (2021) nowcasts gentrification in New York, L.A., and London using Airbnb Data.

Data

The base dataset for the analyses in this paper includes longitudinal survey data from the 2015, 2017, and 2019 waves of the American Housing Survey (AHS).[1] This data is supplemented with information from other sources as described in Table 1.

To gain insight into the socioeconomic and demographic information of housing unit occupants, the analysis dataset collects data from AHS that includes the age, education, marital status, and household income of previous and current residents.[2] Housing characteristics include tenure status, building type, unit size and year that the housing unit was built. This set of characteristics provides basic information about the housing and demographic characteristics that may be predictive of future gentrification.

The AHS data is supplemented with neighborhood characteristics from several sources. In AHS, residents are asked to rate their level of neighborhood satisfaction, a measure that may reflect the respondent's subjective perspective rather than the variation most predictive of gentrification. The analysis dataset therefore instead includes neighborhood characteristics scraped from Redfin, a commercial real estate website that provides a wide range of neighborhood data for each property listing. These neighborhood characteristics include school rating, walking score, environmental risks, and number of amenities

[1] More information on definitions, sampling and non-sampling error, historical changes, and questionnaires is available at < https://www.census.gov/programs-surveys/ahs/tech-documentation/def-errors-changes.html >.
[2] AHS collects both household and person level information in each housing unit. It provides information about the quality and cost of housing in the United States and major metropolitan areas including the physical condition of homes and neighborhoods, the costs of financing and maintaining homes, and the characteristics of people who live in these homes.

including restaurants and parks. Lastly, data from the 2015, 2017 and 2019 waves of ACS are used to create the gentrification indicator.

Table 1. Summary of data sources

Source	Category	Variable
American Housing Survey (2015, 2017, 2019)	Socioeconomic indicators	Household income
		Householder's education attainment
		Householder's age
		Number of young and old children, people
		Family type
		Householder's marriage status
	Housing Characteristics	Building type
		Tenure status
		Garage
		Number of bedrooms, bathrooms
		Condo
		Year built
		Heating fuel source
		Number of renovation jobs done recently
		Unit size
		Rodent
American Community Survey (2015, 2019)[3]	Household income	Household income
Commercial real estate website[4]	Education	Distance to elementary, middle, and high school
	Amenity	Number of restaurants, grocery stores, parks nearby
riskfactor.com and climatecheck.com	Environmental Risks	Flood, storm, drought, heat
Walkscore.com	Walking score	Walk, transit, bike score (out of 100)
GreatSchools.org	Education	School rating (out of 10), Teacher to student ratio

Creating the Gentrification Indicator

The outcome measure for the gentrification models is an indicator variable that identifies which housing units show gentrified between the 2015 baseline survey and the follow-up waves in 2017 and 2019.

[3] The indicators from ACS are aggregated at Census tract level.

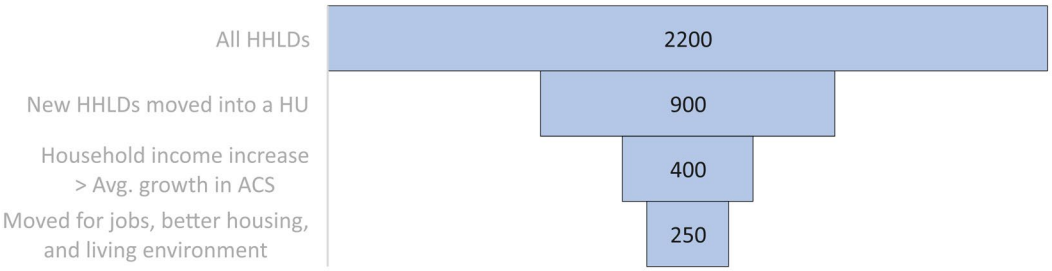
[4] Neighbor information on climate, walkability, and school rating can be obtained from this website.

Specifically, a housing unit is defined to be gentrified if it meets all three of the following criteria: all household members in 2017 or 2019 are different than the members in 2015; household income growth (from 2015 to 2017 or from 2015 to 2019) exceeds the tract level growth rate from ACS; and household members reported moving for better jobs, homes, or neighborhoods.

The first condition indicates whether the housing unit turned over to a new set of residents. The income filter identifies housing units where existing residents were replaced by high-income residents. The final condition requires that high-income earning gentrifiers moved into their housing units for reasons related to jobs, commuting, or a better living environment.[5]

Among 2,200 housing units in 2015, about 41 percent of housing units (900 housing units) had new residents or non- interview cases in either 2017 or 2019, and among the recent movers, 400 housing units reported higher household income in 2017 or 2019.[6] Adding in the third criteria that households recently moved in pursuing a better job and living environment, the gentrification flag yields 250 housing units identified as gentrified between 2015 and 2019.

Figure 1. Creation of Gentrification Flag[7]



Source: U.S. Census Bureau, American Housing Survey, American Community Survey, 2015, 2017, 2019
Note: In accordance with the disclosure avoidance policy, the numerical values presented in the figure have been rounded.

In order to look at the distributions of the gentrification indicator across the Washington, D.C. MSA, the incidence of gentrification is compared by state, metropolitan area, and urban area type. The average share of flagged housing units is 7.91, 10.94, and 14.57 percent in D.C., Maryland, and Virginia respectively. Due to a small sample size in West Virginia, no housing units are identified as gentrified. In terms of metropolitan area type, gentrification occurs more in central city than non-central city (13.15 vs. 11.84 percent). Similarly, gentrified housing units tend to be located in urban cities where the share

[5] Respondents who moved are asked their reason for moving with yes/no questions for 9 potential reasons: forced to move by landlord, bank, government, or disaster; moved because of change in household; moved to reduce commute; moved for family; moved for better home; moved to be in a more desirable neighborhood; moved for job; moved to form household; and other reasons. Appendix A.1 compares the median income of those who answered yes or no to one of the questions for recent movers in 2017. The table demonstrates that movers seeking shorter commuting times, better homes, more desirable neighborhoods, and for job-related reasons tend to earn higher incomes, implying the characteristics of gentrification discussed in the literature.

[6] In accordance with the disclosure avoidance policy, the numerical values presented in the sentence have been rounded.

[7] Non-interview or vacant cases are removed when combined with the questions for recent movers. Out of 2,400 housing units interviewed in Washington D.C. MSA, 2,200 housing units have complete inputs for the gentrification flag in 2017 or 2019.

of gentrification in urbanized areas are 5.01 percentage points higher than the share from rural areas and 3.79 percentage points higher than the share from less urbanized areas.

Table 2. Geographical Distribution of Gentrified Housing Units (weighted)

Geography		Share of Gentrified Housing Units	Number of Gentrified Housing Units	Population
State	Washington D.C.	7.91	18,180	229,800
	Maryland	10.94	87,010	795,500
	Virginia	14.57	134,680	924,600
	West Virginia	0.00	0	20,150
Metropolitan Area Type	Central City	13.15	66,010	502,000
	Non-central city	11.84	173,900	1,468,000
Urban Area Type	Urbanized Area	12.63	224,500	1,778,000
	Urban Cluster	8.85	4,826	54,560
	Rural	7.62	10,510	137,900

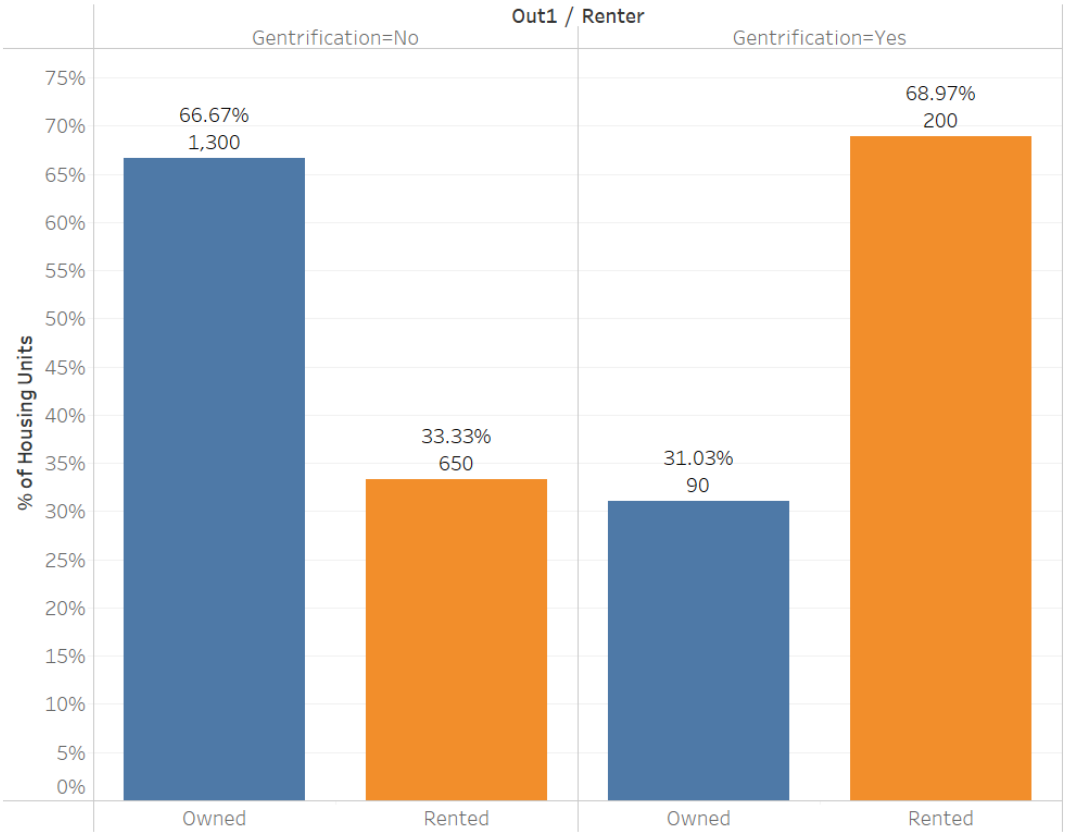
Source: U.S. Census Bureau, American Housing Survey, 2015, 2017, 2019
Note: In accordance with the disclosure avoidance policy, the numerical values presented in the table have been rounded.

Profile of Gentrified Housing Units

This section explores the patterns of gentrification in Washington D.C. MSA by analyzing the selected features by gentrification flag and survey year to identify how each feature increases the probability of gentrification. The distribution of each variable among housing units flagged as gentrified is compared with the non-flagged housing units, considered as non-gentrified. Then, the characteristics of residents that displaced the previous residents surveyed in 2015, defined as “gentrifiers”, are compared with the profile of gentrified residents. Among 56 features selected, 8 features with the highest correlations are selected for this analysis.

Figure 3 shows tenure type by gentrification in 2015. Majority of displaced residents are renters (about 69 percent) whereas two third of non-gentrified residents are homeowners. Although rental properties usually have higher turnover than owner-occupied housing units, income increase between new and old residents are driven by renters who value proximity to job and better neighborhood environment and housing.

Figure 3. Tenure of Gentrified Housing Units by Gentrification Status[8]



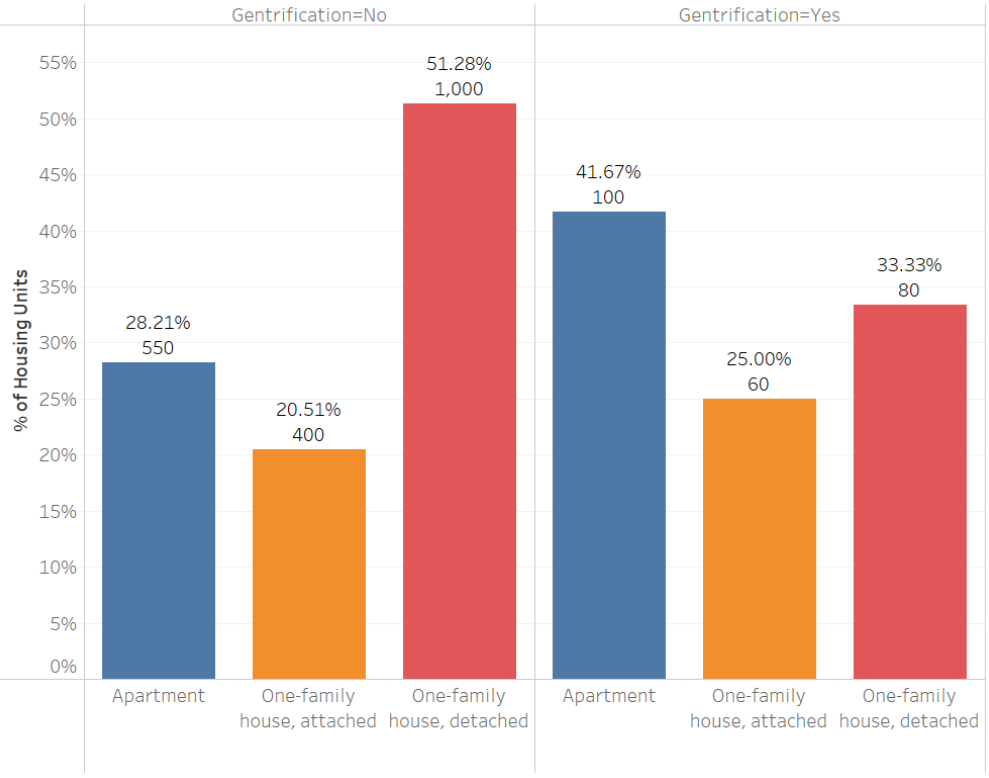
Source: U.S. Census Bureau, American Housing Survey, 2015
Note: Labels include counts and percentages. In accordance with the disclosure avoidance policy, the numerical values presented in the figure have been rounded.

In terms of building type as shown in Figure 4, a greater share of gentrified units are apartments. The share of apartments is 13.46 percentage points higher among gentrified housing units (41.67 vs. 28.21 percent). In 2015 AHS, 51.28 percent of non-gentrified housing units are single-family detached homes whereas this building type accounts for only 33.33 percent of gentrified housing units.

The distributions of building type among gentrified units and gentrifiers shows that gentrifiers in the Washington D.C. MSA are mostly renters living in apartments. This feature may not be generalized to the entire housing market in the metro area due to the non-random attrition of units over time in a panel survey in 2017 and 2019. Also, the sample may not capture single-family houses or multi-family complex newly built after 2015. Despite the potential limitation on external validity, the sample in 2015 is representative of the housing units in Washington D.C. MSA.

[8] Weighted estimates and margin of errors are reported in Appendix A.2.

Figure 4. Building Type of Housing Units by Gentrification Status

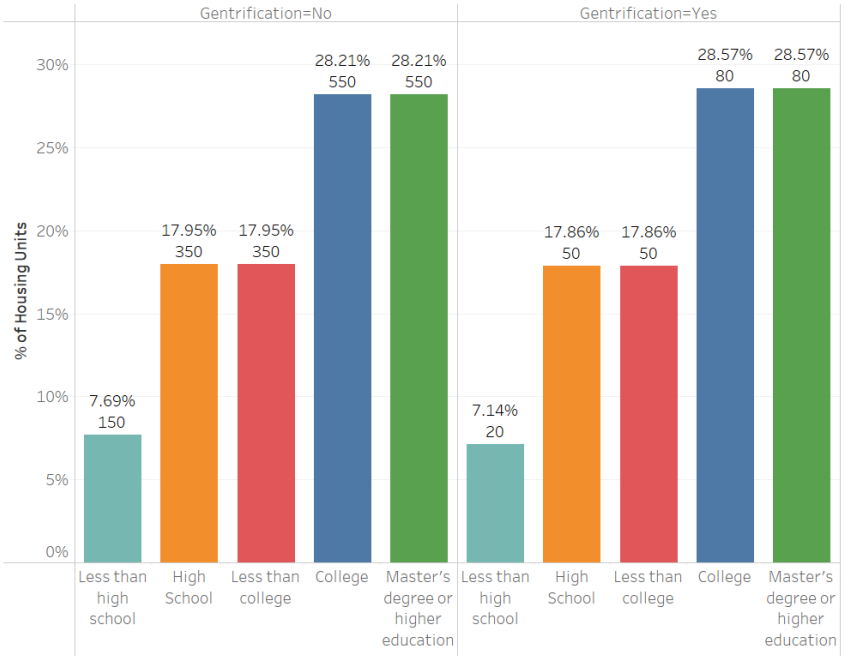


Source: U.S. Census Bureau, American Housing Survey, 2015
Note: Labels include counts and percentages. Note: In accordance with the disclosure avoidance policy, the numerical values presented in the figure have been rounded.

Residents' demographic characteristics in the data include householder's age, race, education, and number of young children under age 6 or old children aged 6 through 17. Conditioned on gentrification flag, there's no correlation between educational attainment and the type of housing unit (Figure 5). However, among those who are gentrified after 2015, less educated householders were mostly displaced by residents with higher education. Since the level of education is highly related to income, less educated residents are more likely to be replaced with educated residents. Figure 6 compares the education level of gentrified householders with the ones of gentrifiers in 2017 and 2019. Householders with less than a college degree were replaced by residents with higher education than the old residents. For residents with at least a college degree, however, most of the new residents were not higher educated than older residents. Among the householders with high school diploma, for example, majority of them were replaced by higher educated householders (about 78 percent) whereas the share of householders among less than college or college degrees displaced by the residents with higher education were relatively lower (about 52 and 44 percent).[9]

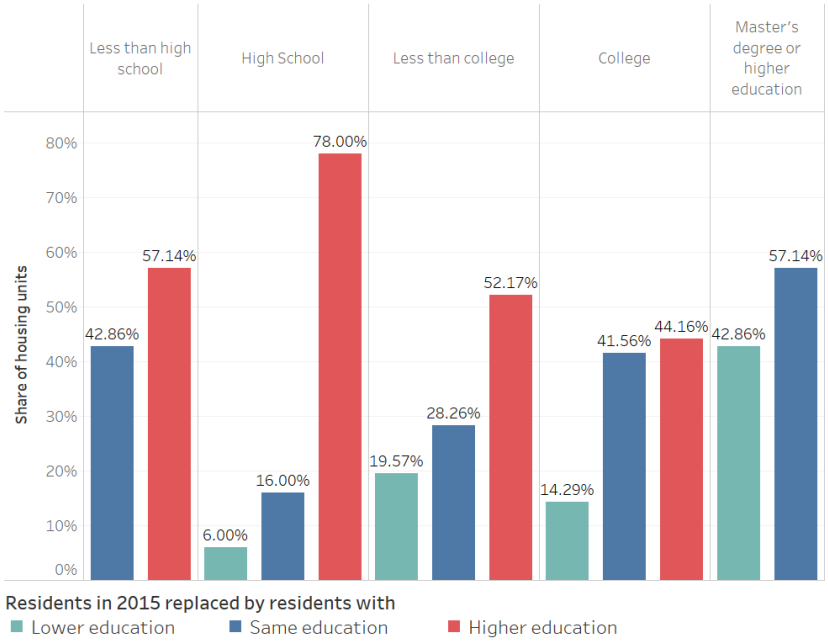
[9] Percentages are calculated based on rounded counts.

Figure 5. Householder's Education Level by Gentrification Status



Source: U.S. Census Bureau, American Housing Survey, 2015
Note: Labels include counts and percentages. In accordance with the disclosure avoidance policy, the numerical values presented in the figure have been rounded.

Figure 6. Comparison of Education Level between Gentrified Householders and Gentrifiers

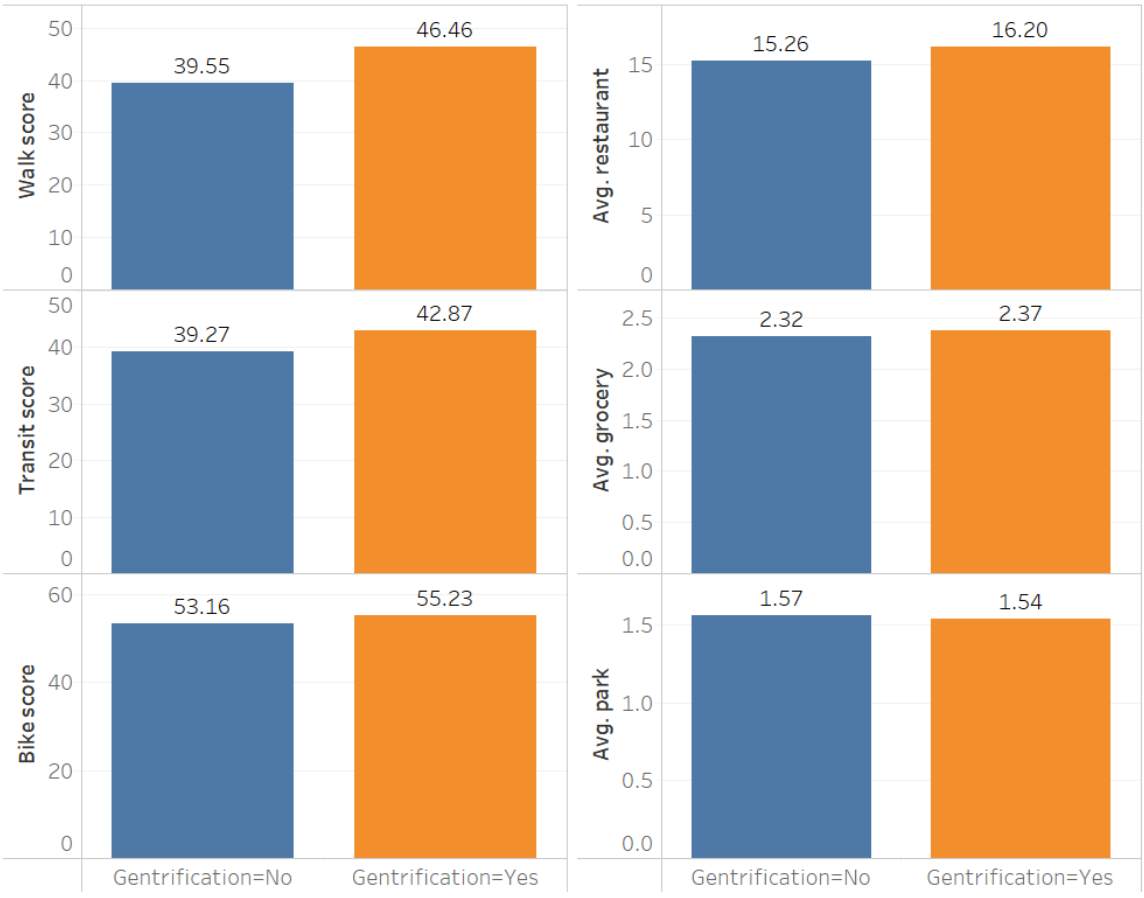


Source: U.S. Census Bureau, American Housing Survey, 2015
Note: Labels include percentages. In accordance with the disclosure avoidance policy, counts are not presented in this figure.

Features representing neighborhood quality obtained from external data sources indicate that housing units in walkable neighborhoods have higher changes of being gentrified. Panel A compares the average walk scores between non-gentrified and gentrified housing units. Walk Score is a walkability index ranged between 0 and 100 based on the distance to amenities such as grocery stores, schools, parks, libraries, restaurants, and coffee shops.[10] The average walk score is about 7 points higher in gentrified housing units and gentrifying neighborhoods tend to have a transit system within a walking distance (39.6 vs 46.5). Bike score is also about 3.6 points higher in gentrified areas indicating a better living environment. Panel B in Figure 7 show the average number of amenity facilities including restaurant, grocery store and park. Excluding the number of parks, gentrified housing units have a greater number of restaurants and grocery stores nearby. The results indicate the gentrified housing units have better amenity and transit accessibility. It is also consistent with the findings on gentrification where gentrifying areas may attract more local business or up-scale establishment (Glaeser et al, 2018).

Figure 7. Average Walk score and Number of Amenity Facilities

Panel A Panel B



Source: External commercial data

[10] Walk Score is a private company that provides walkability services and apartment search tools through a website and mobile applications. Its flagship product is a large-scale, public access walkability index that assigns a numerical walkability score to any address in the United States.

Method

This study tests the performance of six classification models for predicting the gentrification indicator defined in the previous section. Specifically, the analyses estimate models using Logistic Regression (LR), K-nearest Neighbors Classifier (KNN), Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting (GB), all of which are implemented with scikit-learn. Each of these models is estimated using the data on household, housing, and neighborhood characteristics described previously. In total, 56 original features are used and formatted to one-hot vectors.

Logistic regression is a parametric classification model that uses a logistic function to estimate binary output model. It assumes a linear relationship between independent and dependent variables and homoskedasticity of the training data. Also, independent variables are not allowed to be colinear. This model can be used for multiclass classifications and easy to estimate for classification problem. However, it requires proper selection of features and cannot be applied on non-linear classification problems. Random forest model is a classification algorithm consisting of many decisions trees. Decision tree is derived from the independent variables, with a node that has a condition over a feature. Each individual tree in the random forest produces a class prediction and the class with the most votes becomes the model's prediction. Unlike the logistic regression model, random forest model supports non-linearity and addresses collinearity better than LR. The KNN model is a non-parametric method that estimates the probability of a data point to become a member of one group or another based on its similarity. The KNN model is easy to implement and requires few hyperparameters, but the number of clusters should be wisely selected, and it may cost large computation time when a sample size is large. The SVM model is another supervised machine learning model that uses classification algorithms for two-group classification problems. It can be used for non-linear problems and handles outliers well. The Gradient boosting classifiers combine many weak learning models together to create a strong predictive model, using decision trees. The GB model is different from other ML algorithms in that the residual is the gradient of loss function.

Each of these models is estimated using the data on household, housing, and neighborhood characteristics described previously. In total, 56 features are used, and categorical variables are formatted to dummy variables. The data was partitioned using a 70/30 ratio into a training and a test set. The training and test sets contained 1,600 and 650 points respectively.[11] The model was trained using the 2015 input variables to predict the gentrification flag based on 2017 and 2019 data.

Results

The model performance is evaluated based on standard metrics: accuracy, precision, recall and F1 score. Accuracy is a ratio of correctly predicted observations to the total observations. Accuracy can be a good measure when the data sets are symmetric where the number of false positives and false negatives are almost the same. Imbalanced data may overestimate a model's prediction power since most of the accuracy may come from majority classification. Thus, additional metrics need to be considered in addition to accuracy. Precision is the ratio of correctly predicted positive observations to the total

[11] In accordance with the disclosure avoidance policy, the numerical values presented in the sentence have been rounded to the nearest 100 and 50, respectively.

predicted positive observations. Recall is the ratio of precision over accuracy. The F1 score is the weighted average of precision and recall.

Table 3 compares the model performance of every model on each data split during the training stage. The baseline on the first row is estimated using a dummy classifier. Since the dummy classifier does not predict the outcome variable based on input features, other complex ML algorithms are expected to perform better. Except for recall, the machine learning models perform better than the baseline by 23, 22 and 24 percent for accuracy, precision and F1 score respectively. Among machine learning algorithms, the Random Forest Classifier outperforms in all metrics, yielding almost double the base model scores for accuracy and precision. In many ML studies, Random Forest Classifiers have performed well in both classification and regression problems, because the model is non-parametric and does not require much tuning with minimal bias (Breiman, 2001). This model has been also widely used in previous gentrification studies (Reades et al., 2019; Palafox & Monasterio, 2020).

Table 3. Comparison of Predictive Performance

	Base	**Logistic Regression**	**KNeighbors Classifier**	**Random Forest Classifier**	**Support vector machines**	**Gradient Boosting**
Accuracy	0.473	0.718	0.584	0.830	0.625	0.749
Precision	0.472	0.701	0.578	0.809	0.622	0.730
Recall	0.763	0.763	0.625	0.868	0.637	0.790
F1 Score	0.468	0.730	0.600	0.834	0.629	0.759

Source: U.S. Census Bureau, American Housing Survey, 2015

Interpreting the Models: Feature Interpretation

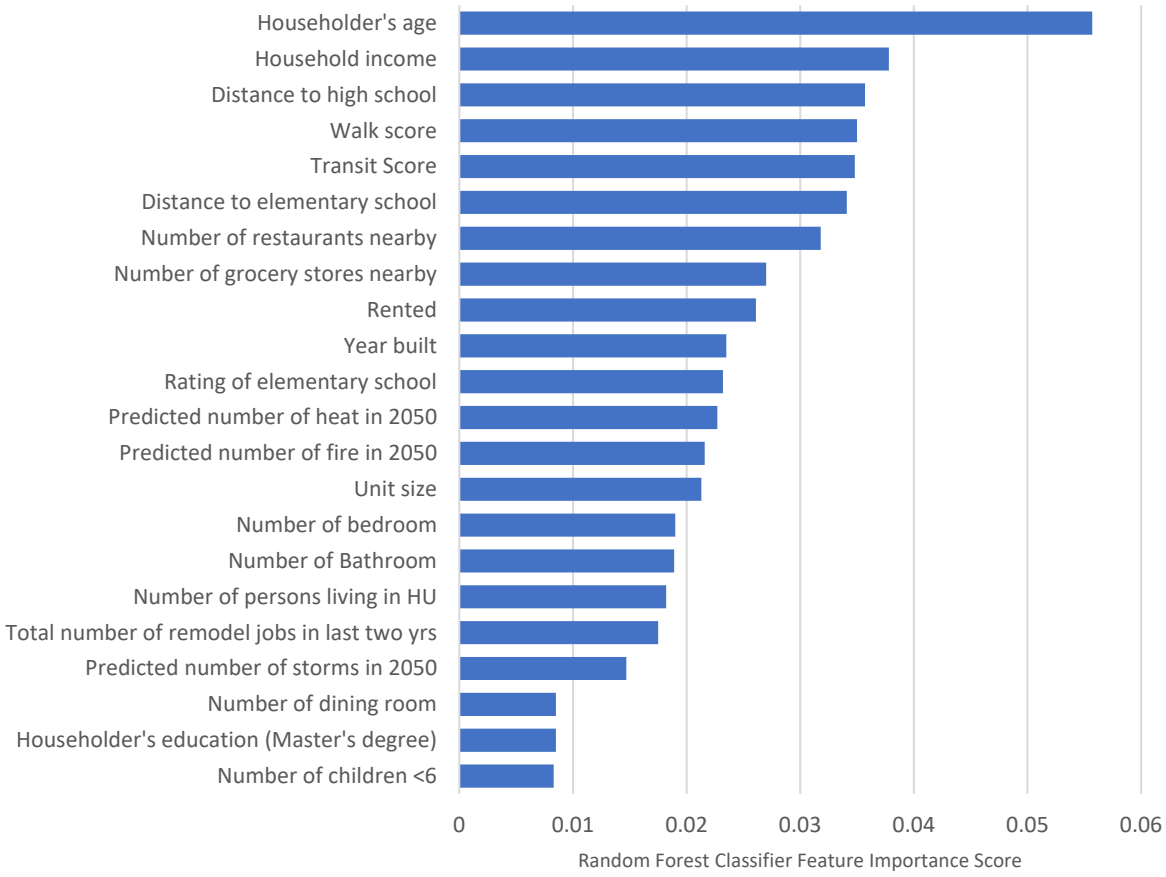
Figure 8 shows the top 22 features that have the highest feature importance in the best performing Random Forest model. Feature importance is a standardized score that calculates the decrease in node impurity weighted by the probability of reaching that node. The node probability is the ratio between the number of observations that reach the node and the total number of samples (Ronaghan, 2018). A higher score of a feature indicates the specific feature has a greater contribution to the model to predict an outcome variable.

Many of the features in the list are the neighborhood attributes from the commercial real estate website, showing the influence of these neighborhood quality measures in predicting gentrification. For example, walk score and transit score are among the 5 most important features. Housing units located in the areas that have shopping centers or a transit system nearby encounter higher risks of gentrification than residences with lower walk/transit scores. Similarly, the number of restaurant and grocery stores affects the decision of high-income earners on their residential location. Environmental factors such as the predicted number of future heat days, fire days, and storms are also ranked as important features in the list.

Similar to the correlation results in the previous section, householder's age, income, and tenure status (of being a renter) are listed as top 10 important features. Number of children aged under 6 also shows strong correlation with the gentrification flag. In terms of housing characteristics, year built is the most important feature, and unit size and the number of bedrooms and bathrooms each show strong

correlation with gentrification flag. Compared to neighborhood quality and the measures of socioeconomic status, these housing characteristics yield relatively smaller importance in predicting gentrification.

Figure 8. Feature Importance: Degree of Usefulness for the Random Forest Model[12]



Source: U.S. Census Bureau, American Housing Survey, 2015

Prediction

As a final step in the analysis, this section illustrates how the preferred model can be applied to the 2019 AHS data to predict future gentrification risk. The previous sections identify the tuned Random Forest classifier as the preferred modeling approach using baseline characteristics from 2015 and a gentrification indicator for 2017 and 2019.[13] The estimates from the Random Forest classifier are then

[12] Feature Importance refers to techniques that measure a score for all the input features for a given model.
[13] For hyperparameter tuning, I used k-fold cross validation (k=7) to tune the maximum depth of tree, the minimum number of samples required to reach a leaf node, the minimum number of samples to split each node, and the number of trees you for taking the averages of predictions.

applied to the characteristics present in the 2019 AHS wave to predict future gentrification in the Washington D.C. area.[14]

Table 4 compares the distribution of gentrification by geography, tenure and building type between 2015 and 2019. The initial rows of the table show the percent of 2019 housing units that the model predicts will be gentrified within 2-4 years based on the preferred random forest classifier. The results show that the predicted MSA level gentrification share using the 2019 data is not statistically different from the estimate for 2015, although it is slightly attenuated. Looking at the gentrification by state and urban areas in both years, D.C. has a about a 10 percentage-point increase in the share of gentrification in 2019 while city areas in Virginia and Maryland have about 4.3 percentage point higher gentrification share in 2019. The gentrification incidence in non-city areas is rather smaller in 2019. City areas in Virginia and Maryland have the highest gentrification share in both 2015 and 2019. Geographically, the gentrified housing units are concentrated in those city areas close to D.C. while some city areas further from D.C. and other rural areas have a scattered pattern of gentrification, yielding smaller gentrification shares.

The bottom half of Table 4 shows the percent of housing units that are renters, separated by the predicted gentrification status. The rental shares among non-gentrified and gentrified units are drastically different, with the rental share at least 25 percentage points higher among units that are predicted to be gentrified in 2019 than among non-gentrified units. The share of apartments among gentrified units also rises by 26.63 percentage points in 2019 from 2015. The large increase in the shares of renters and apartments among gentrified units in 2019 may be explained by the overall increase in the share of rental properties in DMV areas. According to AHS, the estimate for renters in 2019 rose to 37.85 percent from 36.69 percent in 2015 and the share of apartment among renters increased from 66.23 percent to 72.25 percent. As the RF model indicates renters living in an apartment have higher risks of being gentrified, more apartment buildings for lease may yield a large jump in the share of renters and apartments among gentrified units.

Table 4. Patterns of Gentrification by Geography, Tenure, and Building Type (%)

	2015	2019
Share of gentrification by geography		
Washington D.C. MSA	12.18	12.79
District of Columbia	7.91	17.97
City areas in VA and MD	17.57	21.89
Share of renters and apartment by gentrification status in Washington D.C. MSA		
Share of renters		
Gentrified=0	32.55	29.88
Gentrified=1	66.59	92.24
Share of APT		
Gentrified=0	27.25	27.22
Gentrified=1	46.09	72.72

[14] The time gap between the data used for modeling and the inputs used for prediction in 2019 is short enough to apply the estimates from identifying gentrification in 2015. However, the prediction for future gentrification can be limited by the unpredictable dynamics of social and environmental change.

Source: U.S. Census Bureau, American Housing Survey, 2015, 2019
Note: Weights are applied.

Predicted probabilities of gentrification are compared by state in Table 5. Geographically, Washington D.C. has the highest average gentrification chance where residents carry 19 percent of risk of being displaced by higher income earners. When these probabilities are divided into 3 groups -high, medium, and low risk-, Washington D.C. has more high-risk housing units compared to Virginia: the share of high-risk accounts for 18 percent in Washington D.C. whereas about 11 percent of residents are under high risk of gentrification in Maryland. The share of medium-risk housing units is 57.86 and 54.91 percent in Maryland and Virginia, respectively, and the percentage of low-risk housing units ranges between 31.42 to 35.35 percent among the three states.[15]

Table 5. Gentrification Probability by State and Risk Group

	DC	VA	MD
Average gentrification probability	19.20	15.53	14.30
Share of housing units with			
High risk (Prob>=0.35)	17.97	13.37	10.72
Medium risk (0.15<=Prob<0.35)	46.69	54.91	57.86
Low risk (Prob <0.15)	35.35	31.72	31.42

Source: U.S. Census Bureau, American Housing Survey, 2019
Note: Weights are applied.

Looking at the geographical distribution of predicted gentrification cases, geographically these housing units are clustered in specific location. To highlight this clustering feature, the average distance between units flagged as gentrified is calculated for each ZIP code. Then, the average distance of all units is computed for each ZIP code to obtain the ratio between the average distance of flagged units and all units. Any ZIP codes with the ratio smaller than 0.8 are flagged as the areas with clusters of high-risk housing units.

Table 6 shows the number of flagged ZIP codes by state as well as the average ratios of selected ZIP codes. Two thirds of the ZIP codes in D.C. have the clusters of gentrification cases where 7 out of 11 ZIP codes are selected. In Virginia, the clusters occur in about half of all ZIP codes where 31 ZIP codes are identified as clustered areas among 62 ZIP code areas.

Geographically, most of the gentrification takes place near the Capital Beltway, and the high-risk housing units are clustered in areas that have seen significant real estate development. These are the areas that have experienced major redevelopment over the past decade (Orfield, M. W., 2019). In addressing ongoing housing affordability issues, county and city officials have approved large scale development and zoning applications and development projects to build commercial stores and rental apartment buildings. Also, to provide housing assistance to low-income families, housing assistance programs such as public housing, low-income housing tax credits, and housing vouchers have been offered. Despite such efforts, the increasing influx of highly educated young adults in Washington D.C.

[15] High, medium, and low risk groups are divided by two cutoff probabilities: 0.35 for high and medium-risk groups and 0.15 for medium and low-risk groups.

metro area has already taken over housing supply in newly developed areas, accelerating the gentrification among low-income residents (Kingsley, 2017).

Table 6. Average distance ratio and number of zip codes with high-risk housing units clustered

State	Average of ratio	Number of clustered areas	Total number of zip codes
D.C.	0.65	7	11
MD	0.55	17	40
VA	0.63	31	62

Source: U.S. Census Bureau, American Housing Survey, 2019

Discussion

Despite the abundance of policy discussions on home affordability and gentrification issues in D.C. areas, there hasn't been enough empirical studies focusing on this area compared to the research on other metropolitan areas such as New York, L.A., or San Francisco. Washington D.C. MSA is considered as one of the most expensive areas to purchase a house or pay housing rents. The case study in D.C. metropolitan areas shows the patterns of gentrification that gentrifiers are largely concentrated in high-amenity neighborhoods. The flagged units are located in high-income areas, indicating the data and the model pick up the areas where gentrification has already continued.

The findings are important evidence for policy makers in nowcasting gentrification. By applying the machine learning technique models to newly collected survey data, we can identify which areas are clustered with high-risk housing units and then more targeted measures for low-income residents can be implemented before private developers already materialize the gentrification.

Furthermore, this study suggests a potential solution for constructing the early warning systems. Departing from traditionally used models to predict gentrification status, Random Forest classification model produces more accurate classification prediction. Existing studies emphasize the lack of early warning systems that predict which housing units will have a higher gentrification risk. Constructing early warning systems should be an effective measure for gentrification since private investors have already identified demand for housing and high-quality neighborhood and actualized the gentrification prediction (Rittenbruch et al., 2021; Thackway et al, 2021).

Methodologically, incorporating external data source significantly improves prediction and complements Census survey data. This study provides another example that external neighborhood information has the biggest contribution to predicting gentrified units. The metrics for model performance drastically increases when external data is included compared to using AHS data only.

Conclusion

Although many gentrification studies explore the consequences and the patterns of neighborhood and residents' change retrospectively, studies on nowcasting gentrification have been limited. Overcoming the restrictions from traditional regression models and survey data which often lacks detailed neighborhood information, this paper identifies gentrifying housing units by incorporating external data source and applying machine learning algorithms.

Using longitudinal aspect of AHS data, the study detects housing units with new residents with much higher income and finetunes the gentrification cases adding more filters with questions about the reasons to move. After implementing several ML models, Random Forest Classifier model is adopted for the best prediction model which yields 83 percent accuracy, 81 percent prediction, and 87 percent recall score. Further analysis on the profile of gentrifiers and feature importance suggests that highly educated young adults looking for apartments in walkable urban neighborhoods have driven the gentrification in Washington D.C. MSA. Using a machine learning model that predicts gentrification instead of responding to what already happened would provide more opportunities for policymakers to design solutions for gentrifying housing units.

Appendix

A.1. Median Income among recent movers in 2017

MOVFORCE: was forced to move by landlord, bank, government, or disaster	N	Median
Yes	30	60,000
No	600	77,000
RMCHANGE: moved because of change in household	N	Median
Yes	100	74,500
No	450	78,000
RMCOMMUTE: moved to reduce commute	N	Median
Yes	100	85,000
No	500	75,000
RMFAMILY: moved for family	N	Median
Yes	100	55,000
No	500	80,000
RMHOME: moved for better home	N	Median
Yes	250	95,000
No	350	67,500
RMHOOD: moved to be in more desirable neighborhood	N	Median
Yes	200	81,500
No	400	72,000
RMJOB: moved for job	N	Median
Yes	150	95,000
No	450	74,000
RMOWNHH: moved to form household	N	Median
Yes	200	75,000
No	400	78,000
RMOTHER: another reason	N	Median
Yes	100	59,000
No	450	83,300

Source: U.S. Census Bureau, American Housing Survey, 2017

Note: In accordance with the disclosure avoidance policy, the numerical values presented in the table have been rounded.

A.2. Statistical test results for Figure 3

Share of renters vs. homeowners	Difference	Standard Error	T statistics
Gentrified units in 2015	0.331	0.042	7.954

Source: U.S. Census Bureau, American Housing Survey, 2015

A.3. Statistical test results for Figure 4

Share difference test between gentrified and non-gentrified units	Year	Difference	Standard Error	T statistics
Apartment	2015	0.1761	0.033	5.379
Single-family detached homes		-0.1908	0.037	-5.106

Source: U.S. Census Bureau, American Housing Survey, 2015

A.4. Statistical test results for Figure 5

Share difference test between gentrified and non-gentrified units	Year	Difference	Standard Error	T statistics
College	2015	0.006	0.040	0.155
Master's degree or higher education		0.008	0.040	0.194
High school		0.014	0.050	0.277
Less than college		-0.029	0.051	-0.573
less than high school		0.002	0.076	0.021

Source: U.S. Census Bureau, American Housing Survey, 2015

A.5. Statistical test results for Figure 6

Share of housing units replaced by higher education	Year	Difference	Standard Error	T statistics
Less than high school vs Less than college	2015, 2017, 2019	0.049	0.084	0.593
High school vs College		0.338	0.06	5.542
High school vs Less than college		0.258	0.069	3.732
Less than high school vs College		0.13	0.079	1.646

Source: U.S. Census Bureau, American Housing Survey, 2015, 2017, 2019

Reference

Alejandro, Y., & Palafox, L. (2019, October). Gentrification prediction using machine learning. In *Mexican International Conference on Artificial Intelligence* (pp. 187-199). Springer, Cham.

Atkinson, R. (2000). Measuring gentrification and displacement in Greater London. *Urban studies*, *37*(1), 149-165.

Barton, M. (2016). An exploration of the importance of the strategy used to identify gentrification. *Urban Studies*, *53*(1), 92-111.

Beauregard, R. A. (2013). The chaos and complexity of gentrification. In *Gentrification of the City* (pp. 51-71). Routledge.

Binder, A. J., & Bound, J. (2019). The declining labor market prospects of less-educated men. *Journal of Economic Perspectives*, *33*(2), 163-90.

Bostic, R. W., & Martin, R. W. (2003). Black home-owners as a gentrifying force? Neighbourhood dynamics in the context of minority home-ownership. *Urban Studies*, *40*(12), 2427-2449.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brummet, Q., & Reed, D. (2019). The effects of gentrification on the well-being and opportunity of original resident adults and children.

Byrne, J. P. (2002). Two cheers for gentrification. *Howard LJ*, *46*, 405.

Chapple, K., & Zuk, M. (2016). Forewarned: The use of neighborhood early warning systems for gentrification and displacement. *Cityscape*, *18*(3), 109-130.

Ding, L., & Hwang, J. (2020). Effects of gentrification on homeowners: Evidence from a natural experiment. *Regional Science and Urban Economics*, *83*, 103536.

Ding, L., Hwang, J., & Divringi, E. (2016). Gentrification and residential mobility in Philadelphia. *Regional science and urban economics*, *61*, 38-51.

Easton, S., Lees, L., Hubbard, P., & Tate, N. (2020). Measuring and mapping displacement: The problem of quantification in the battle against gentrification. *Urban studies*, *57*(2), 286-306.

Ellen, I. G., & O'Regan, K. M. (2011). How low income neighborhoods change: Entry, exit, and enhancement. *Regional Science and Urban Economics*, *41*(2), 89-97.

Freeman, L. (2005). Displacement or succession? Residential mobility in gentrifying neighborhoods. *Urban Affairs Review*, *40*(4), 463-491.

Glaeser, E. L., Kim, H., & Luca, M. (2018, May). Nowcasting gentrification: using yelp data to quantify neighborhood change. In *AEA Papers and Proceedings* (Vol. 108, pp. 77-82).

Glaeser, E. L., Kominers, S. D., Luca, M., & Naik, N. (2018). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, *56*(1), 114-137.

Helms, A. C. (2003). Understanding gentrification: an empirical analysis of the determinants of urban housing renovation. *Journal of urban economics*, *54*(3), 474-498.

US Department of Housing and Urban Development (2018), "Displacement of Lower-Income Families in Urban Areas Report", HUD, Office of Policy Development and Research

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.

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.

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.

Kingsley, G. T. (2017). Trends in housing problems and federal housing assistance. *Washington, DC: Urban Institute*.

Lees, L., Slater, T., & Wyly, E. (2013). *Gentrification*. Routledge.

Mayer, N. S. (1981). Rehabilitation decisions in rental housing: an empirical analysis. *Journal of Urban Economics*, *10*(1), 76-94.

McKinnish, T., Walsh, R., & White, T. K. (2010). Who gentrifies low-income neighborhoods?. *Journal of urban economics*, *67*(2), 180-193.

Melchert, D., & Naroff, J. L. (1987). Central city revitalization: A predictive model. *Real Estate Economics*, *15*(1), 664-683.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.

Ogutu, J. O., Piepho, H. P., & Schulz-Streeck, T. (2011, December). A comparison of random forests, boosting and support vector machines for genomic selection. In *BMC proceedings* (Vol. 5, No. 3, pp. 1-5). BioMed Central.

Orfield, M. W. (2019). *American Neighborhood Change in the 21st Century*. https://www.law.umn.edu/sites/law.umn.edu/files/metro-files/american_neighborhood_change_in_the_21st_century_-_full_report_-_4-1-2019.pdf

O'Sullivan, A. (2005). Gentrification and crime. *Journal of urban economics*, *57*(1), 73-85.

Palafox, L., & Ortiz-Monasterio, P. (2020, July). Predicting gentrification in Mexico city using neural networks. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-5). IEEE.

Prince, S. (2016). *African Americans and gentrification in Washington, DC: Race, class and social justice in the nation's capital*. Routledge.

Reades, J., De Souza, J., & Hubbard, P. (2019). Understanding urban gentrification through machine learning. *Urban Studies*, *56*(5), 922-942.

Redfern, P. A. (2003). What makes gentrification'gentrification'?. *Urban studies*, *40*(12), 2351-2366.

Ronaghan, S. (2018, May 11). The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark? *Towards Data Science.* https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

Shin, H. B., Lees, L., & López-Morales, E. (2016). Introduction: Locating gentrification in the global east. *Urban Studies*, *53*(3), 455-470.

Thackway, W., Ng, M. K. M., Lee, C. L., & Pettit, C. (2021). Building a predictive machine learning model of gentrification in Sydney.

Vigdor, J. L., Massey, D. S., & Rivlin, A. M. (2002). Does gentrification harm the poor?[with Comments]. *Brookings-Wharton papers on urban affairs*, 133-182.

Ye, T., Johnson, R., Fu, S., Copeny, J., Donnelly, B., Freeman, A., ... & Ghani, R. (2019, July). Using machine learning to help vulnerable tenants in new york city. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies* (pp. 248-258).

Zheng, D., Hu, T., You, Q., Kautz, H., & Luo, J. (2015). Towards lifestyle understanding: Predicting home and vacation locations from user's online photo collections. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 9, No. 1, pp. 553-560).