Using Gentrification Data to Address Systematic Differences Between the New York and San Francisco MSA

Citadel Europe Regional Data Open Fall 2020

Team 10

Nikodem Czarlinski, Karan Sharma, Aaron Yuan, Jeevan Singh Bhoot

Trinity College, University of Cambridge

October 24, 2020

1 Non-technical Report

Gentrification is the process whereby the character of a poor urban area is changed by wealthier people moving in, often displacing current inhabitants. Analysis of data from the New York Metropolitan Statistical Area (MSA) census revealed a uniquely high correlation between the eligibility of an area to soon undergo gentrification, and the number of Hispanic Latin people and the number of African-Americans living in these areas. To further study this relationship, we ask questions of whether there are potential links between race and the rate of gentrification from 2010-2018.

For our investigation, we used the 'random forests' algorithm to classify areas on the basis of their gentrification, which yielded 99%+ accuracy. We then applied this model outside the New York MSA, on the San Francisco MSA. By comparing our predictions to the actual rates of gentrification in the SF MSA, we found that our model's accuracy was less than before - 87% accuracy rate. In order to explain this, we researched systematic differences between the two regions, including the differences in house prices, cost of living, and incomes.

2 Technical Report

2.1 Methodology

2.1.1 Measuring Gentrification

The concept of gentrification is abstract, and a bit subjective. There is no single intrinsic property of a tract that tells the tale of the gentrification that is or is not taking place in the tract. Many have tried in the past to quantify gentrification, especially Prof. Lance Freeman. In 2005, he published a paper detailing a criterion that tracts would have to pass in order to be 'eligible for gentrification', and a second criterion to identify tracts that had gentrified over a given time period. This definition has been further improved by the Governing Magazine in their 2015 'national report on gentrification', and we will be using their criteria in our study. More on this in 2.1.3 Definitions.

2.1.2 Data Manipulation

While browsing the census data of the New York MSA provided by Citadel, we found anomalies in some of the entries - extremely negative values for home values, household incomes and 0 values for populations.

Initially, we thought these values may have been overflow errors, however after opening the data in several different applications, we were unable to retrieve the 'correct data'. We came to the conclusion that these must have been fundamentally erroneous entries in the table, so to clean the data we removed all rows containing such entries. The removal of these corrupt tracts were also performed on all other data-sets.

Other modifications/additions to the data are listed below:

- Several new columns added to describe the proportions of the populations of tracts classed by ethnicity were added to the existing tracts data for regression.
- Educational data from 2010 and 2018 for the number of bachelor's degree holders per tract was added from the original census data to the tracts included in our study. (Not provided by Citadel)

- Occupational data classed by sector was added to each tract for the years 2010 and 2018 with all years inclusive. (Not provided by Citadel)
- Two new Boolean columns were also added to the data-set. These corresponded to whether or not the tract in question passed the gentrification eligibility test and the actual gentrification test. 1 for pass and 0 for fail.

2.1.3 Definitions

Following on our use of the Governing Magazine's criteria, the definition of the criterion for gentrification (criterion 2) were misleading. Their website quotes that "an increase in a tract's educational attainment, as measured by the percentage of residents age 25 and over holding bachelor's degrees, was in the top third percentile of all tracts within a metro area" must be achieved by a tract for it to pass this point of the criterion. This language suggests that a tract must move from being relatively poorly educated to being the top 3% of improvements. The 3% figure is very extreme, in fact using this criteria we calculated the number of gentrified tracts was just 1 between the years of 2010 and 2018, which is too low to provide proper statistical analysis. This suggested that something was wrong...

The Wikipedia page on gentrification quotes the Governing Magazine's report from 2015 with the language correction:

"the percentage of increase in home values in the tract was in the top 33rd percentile when compared to the increase in other census tracts in the urban area"

As expected, this more reasonable criterion provided a much more realistic result of 36 tracts having passed both criteria for eligibility and gentrification.

After this correction, we split the criteria for gentrification into two tests, one for eligibility and one for gentrification. This ensured that our test for actual gentrification, where the metrics tested for general improvement, only tested among the set of eligible tracts. The amended criteria for both tests with explanations of which subsets of the data were used are listed below.

The criteria for test one were:

- The tract had a population of at least **500** residents within the time frame 2010-2018 and was located within a central city.
- The tract's median household income was in the **bottom 40th percentile** when compared to all tracts within its metro area at the beginning of the decade.
- The tract's median home value was in the **bottom 40th percentile** when compared to all tracts within its metro area at the beginning of the decade.

For test two:

- An increase in a tract's educational attainment, as measured by the percentage of residents age 25 and over holding bachelor's degrees, was in the **top third** of all tracts within a metro area. *We assumed that this was the top third of increases, not just changes.*
- A tract's median home value **increased** when adjusted for inflation.

- The percentage increase in a tract's inflation-adjusted median home value was in the **top third** of all tracts with a metro area. *We again assumed that this was the top third of increases, not just changes.*

Tracts passing both of these tests would qualify as having gentrified in the time period being measured, i.e 2010-2018.

2.1.4 Data Awry, Spotted by Eye

This section is dedicated to our experiences when performing sanity checks on our results and then implementing fixes to our methodology.

- **Obtaining only 1 gentrified tract.** As specified in the 'definitions' section, our first implementation of both sets of criteria for gentrification eligibility and actual gentrification returned only 1 tract. Our method was then extensively checked before challenging the definition given to us by the Governing website. Further research into the paper that the article by Governing was based on revealed that Governing had made a stylistic typo and so we replaced the criteria from the top 3% to the top third (33%) of the distribution of educational attainment improvement and median home value increases. Our modified definition of this criteria led to 13 tracts having passed test 1 and test 2, a much more realistic result.
- **Increases! Not changes...** This was specific for test 2. The criteria that we have now defined for both the education and median home value is based on the sample of tracts that showed improvement, or increases, in these metrics. Previously, we were calculating percentile values for change, which could be positive or negative. This resulted in the percentile values being negative, which we noticed as an undesired result, prompting us to look back into where we made the mistake. We were able to rectify this problem.

2.1.5 Assumptions

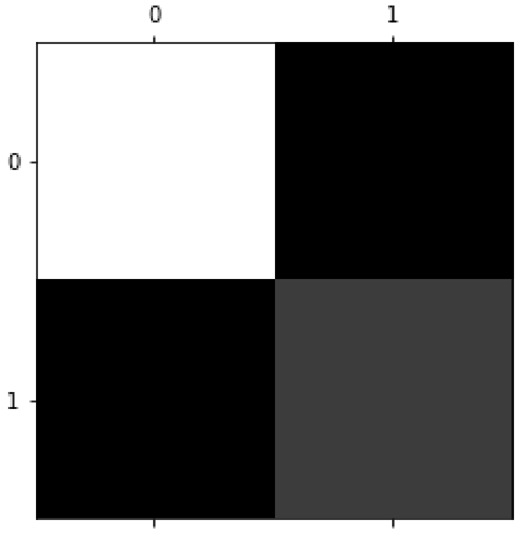
In our study, we made the following assumptions:

- We assumed that all census' data we used were representative of the populations they were drawn from.
- We assumed that the criteria described in the Governing Magazine's report accurately quantifies eligibility for gentrification and gentrification in the years 2010-2018.

2.1.6 Modelling

Overall, we trained three different predictive models. The first was to predict the gentrification eligibility (test one) of tracts in both the NY-MSA and SF-MSA, the second was to predict the actual gentrification rate (test 2) of these areas, and the third was to predict these tracts' improvements using metrics in incomes, house prices and educational improvements. These were all optimised in their hyper-parameters which will be included in tables under them. During the training process, we performed a randomised search with 3 fold cross validation and 500 iterations to find the best set of hyper-parameters, using a scoring system of the f1 value. The models were also all trained and tested using independent subsets of data from the total census in order to avoid over-fitting.

We trained a random forest classifier model using subsets of New York MSA data in 2010 and 2018 to predict gentrification eligibility (test 1). A random forest is an ensemble of decision trees where each classifier makes a prediction and the overall prediction is based on the most common vote. This is also a process known as hard voting. This is shown by the following confusion matrix, where the 0 indicates negative and 1 indicates positive. The rows show the predictive model results and the columns show the actual results. The perfect result of 100% accuracy would be for the main diagonal to be completely white and the other cells to be black (0 absolute values). The hyper-parameters are also listed in a table underneath.

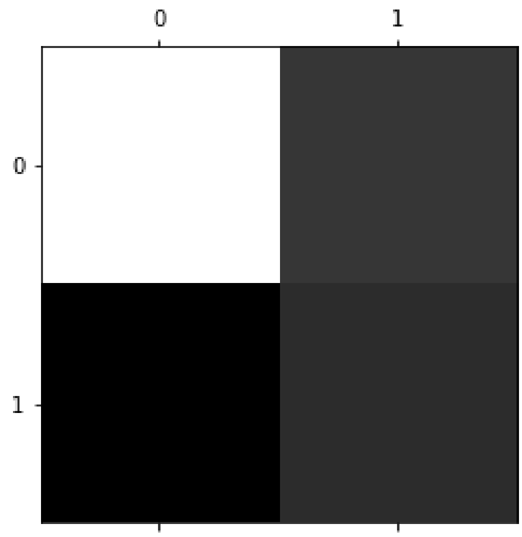


$$\begin{bmatrix} 3638 & 3 \\ 2 & 855 \end{bmatrix}$$

f1 score = 0.9971, precision = 0.9965, recall = 0.9977

Hyper-parameters
n_estimators = 1823 min_samples_split = 4 min_sample_leaf = 1 max_features = auto max_depth = 76 class_weight = None bootstrap = False

This model for predicting gentrification eligibility had an accuracy rate of 99.65%. Using the same model, we also predicted the gentrification eligibility rates using test 1 of the San Francisco with an 84.64% accuracy rate described by the confusion matrix below.

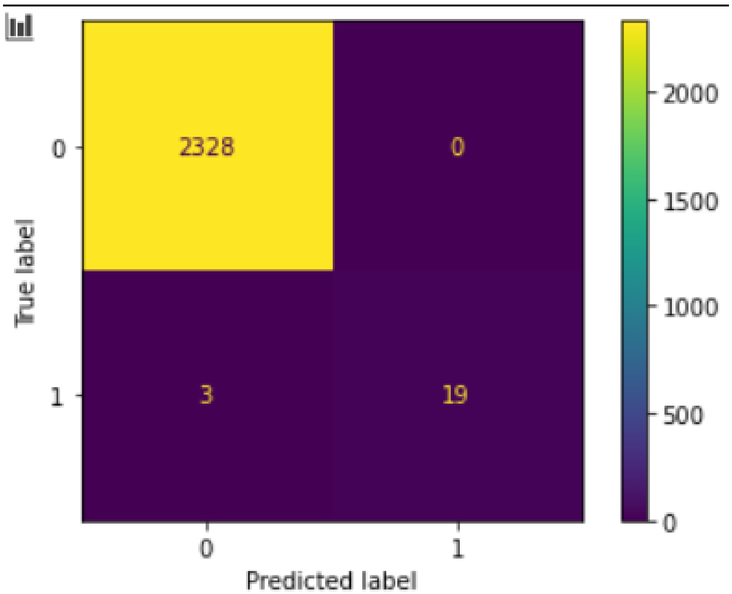


$$\begin{bmatrix} 6171 & 1321 \\ 0 & 1106 \end{bmatrix}$$

f1 = 0.626, precision = 0.4557, recall = 1

Note that the false negatives were perfectly black at 0 score but the number of false positives were relatively high, signalling systematic racial differences between the regions of the San Francisco MSA and the New York MSA based on income and housing quality metrics used in test 1.

The second model predicts the rate of actual gentrification. For this, we trained a model using 2010-2018 data, all years inclusive, to perform both test 1 and test 2. Then, when testing it on NY MSA data, making sure to use independent clusters of subsets of data from the census, the following confusion matrix was obtained.



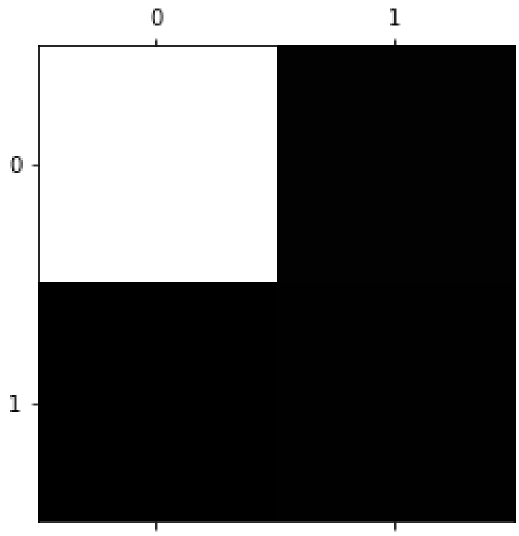
$$\begin{bmatrix} 2328 & 0 \\ 3 & 19 \end{bmatrix}$$

f1 = 0.925373, precision = 0.861111, recall = 1

Hyper-parameters
n_estimators = 10
min_samples_split = 2
min_sample_leaf = 3
max_features = auto
max_depth = 20
class_weight = balanced_subsample
bootstrap = True

What is interesting about this model is that even after this model is then trained only on the variables of races, it is still able to predict the gentrification rate using test 1 and test 2 in the NY MSA. When ranking the features of the different races, the group that was found to be most disproportionately affected by gentrification were the Hispanic-Latin group, closely followed by the Black/African-American group.

The third model was based on tract improvement as opposed to gentrification. Any tract that would have qualified for gentrification using test 2 criteria but did not pass test 1 criteria for gentrification eligibility were included as having passed overall. The model had the following confusion matrix below when testing against independent subsets of NY MSA data. The optimised hyperparameters are also shown.

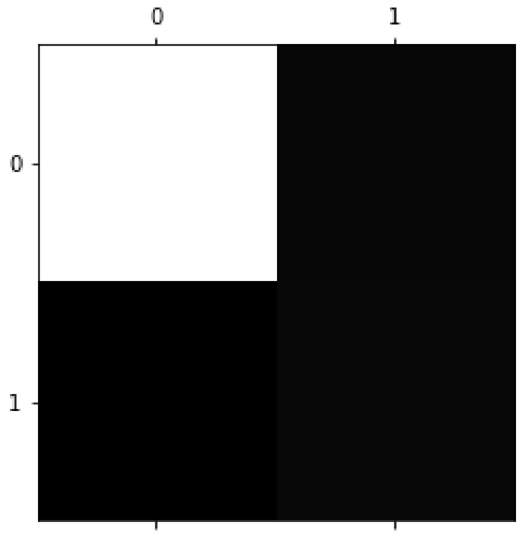


$$\begin{bmatrix} 4294 & 39 \\ 2 & 35 \end{bmatrix}$$

f1 value = 0.63063, precision = 0.47297, recall = 0.94595

Hyper-parameters
n_estimators = 10
min_samples_split = 2
min_sample_leaf = 3
max_features = sqrt
max_depth = 10
class_weight = balanced
bootstrap = False

Given that the construction of this model was based on looser constraints on the criteria for test 2, then this was expected. However, it is still favourable to observe such high predictive capabilities for an improvement model. Again, the model was trained using clustered subsets of the data that were independent of the data that was used to test the model, therefore reducing the chance of the model over-fitting. We then used this model on San Francisco MSA data to produce the following confusion matrix, yielding a 97.164% accuracy rate.



$$\begin{bmatrix} 896 & 27 \\ 0 & 29 \end{bmatrix}$$

f1 score = 0.6824, precision = 0.5179, recall = 1

2.2 Analysis

As a preliminary investigation, we carried out multi-linear regression on gentrification eligibility and several factors listed below, using data from 2010, and 2010-2018 (all years inclusive). The table of regression values is shown below.

Variable	r-value (2010)	r-value (2010-2018)
gentrify_elig	1.000000	1.000000
no_hisp_latin	0.325739	0.330896
no_nonh_blacks/aas	0.262827	0.242014
no_amerinds_alskns	0.056961	0.074572
no_nonh_others	0.041292	0.041577
population	0.024160	0.021747
no_nonh_hawaii_pacific	-0.000848	0.013647
tract	-0.027139	-0.021100
no_nonh_multi	-0.028295	-0.045377
no_nonh_asians	-0.090139	-0.104720
state	-0.132786	-0.167929
county	-0.210060	-0.211035
no_nonh_caucasians	-0.296306	-0.287119
household_income	-0.440283	-0.441597
home_value	-0.488391	-0.439094

In terms of relative values, we can clearly see a positive correlation between the eligibility of gentrification with specific race groups such as Hispanic-Latins and African-Americans (non-Hispanic blacks). These were contrasted with the relatively largely negative regression values for the number of non-Hispanic Caucasians. Note that the values for Asian households were weakly negative.

By then adding features to describe the proportionate sizes of these groups within their tracts, as well as variables corresponding to the proportion of non-whites, we obtained the following regression table values.

Variable	r-value (2010)	r-value (2010-2018)
gentrify_elig	1.000000	1.000000
proportion_non_white	0.364748	0.350848
proportion_hisp_latin	0.349615	0.364114
proportion_nonh_black	0.262827	0.231117
proportion_amerinds_alskns	0.056961	0.059358
proportion_nonh_others	0.024284	0.018607
proportion_nonh_hawaii_pacific	-0.007678	0.007229
proportion_nonh_multi	-0.040574	-0.060249
proportion_nonh_asians	-0.126842	-0.142732
proportion_nonh_caucasians	-0.364748	-0.350848

Observing the racial variables, we can see polarising extremes between the r-values for the proportion of non-whites and the proportion of non-Hispanic Caucasians. As the criteria for eligibility of gentrification are proxy to indicators for inequality in incomes and housing quality, this data is an insight into whether or not there is systematic discrimination based on race in the New York MSA.

We then used 2010 San Francisco MSA data to find if these trends in New York were more universal across the nation. This would also help us find if our model for New York was over-fitted, which would present itself as any large discrepancies in the prediction accuracy and precision rates. Our findings from the regression values are listed below.

Variable	Regression Value
gentrify_elig	1.000000
proportion_nonh_black	0.510570
proportion_hisp_latin	0.494656
no_nonh_blacks/aas	0.479871
proportion_non_white	0.462821
no_hisp_latin	0.397834
no_non_white	0.271602
proportion_amerinds_alskns	0.204831
no_amerinds_alskns	0.182181
proportion_nonh_hawaii_pacific	0.137445
no_nonh_hawaii_pacific	0.118210
proportion_nonh_others	0.002359
no_nonh_others	-0.007769
proportion_nonh_multi	-0.086612
no_nonh_multi	-0.102543
proportion_nonh_asians	-0.158762
no_nonh_asians	-0.175580
no_nonh_caucasians	-0.405901
proportion_nonh_caucasians	-0.462821

We observe from these regression values that the proportion of non-Hispanic blacks was the highest correlative variable with the gentrification eligibility in San Francisco, with the proportion of Hispanic-Latins following closely behind. Note that these values are consistently much higher than for New York. Although hard to explain, this suggests a larger racial inequity in living standards in the San Francisco MSA than the New York MSA.

To further try to explain this, we then utilised census data involving the occupations and education held by households of New York and San Francisco in 2010 to potentially describe any underlying systematic differences that could relate to race. The regression values are summarised in the table below for New York. Note that these values are still from multiple regression against gentrification eligibility from test 1.

Variable	Regression Value
gentrify_elig	1.000000
2010_arts_recreation_accom	0.043675
2010_armed_forces	0.027868
2010_construction	0.010709
2010_retail	-0.000540
2010_edu_health_social	-0.058178
2010_public_admin	-0.069780
2010_female<25_bachelors	-0.102783
2010_male<25_bachelors	-0.142074
2010_information	-0.172911
2010_finance	-0.188538
2010_bachelors	-0.236275
2010_>25bachelors	-0.240077
2010_percent_>25bachelors	-0.338479

The trend of negative correlations between the eligibility of gentrification and the metrics with education are to be expected, since the construction of our criteria for test 1 was largely negatively correlated with the household incomes. Since holding a bachelor's degree generally increases the chance for the individual to attain higher paid skilled jobs, then this result could be simply explained by this relationship.

To solidify these claims, we took every single feature that was both featured in the original census data-set as well as the features that we added ourselves. The following table has the notable relationships that will support a story of disproportionate experiences of ethnic minorities, specifically Hispanics and Blacks, in gentrification.

Variable 1	Variable 2	Regression Value
proportion_hisp_latin	proportion_bachelors	-0.464235
proportion_hisp_latin	2010_finance	0.433671474008472

2.3 Conclusion