There Goes the Neighborhood

Predicting Gentrification and the Socioeconomic Impact Income has on U.S. Counties

Dominique Reynolds
Filsan Yousuf
Melynda Schreiber
Naomi Rankin
Omar Pineda Jr

Overview

- Problem and Solution
- Dataset
- Exploratory Data Analysis
- Model
- Conclusion

What is Gentrification?

Gentrification is the process in which a lower-income area experiences an influx of middle-class or wealthy people who renovate and rebuild homes and businesses. This often results in an increase in property values and the displacement of earlier, usually lower-income residents.

Problem

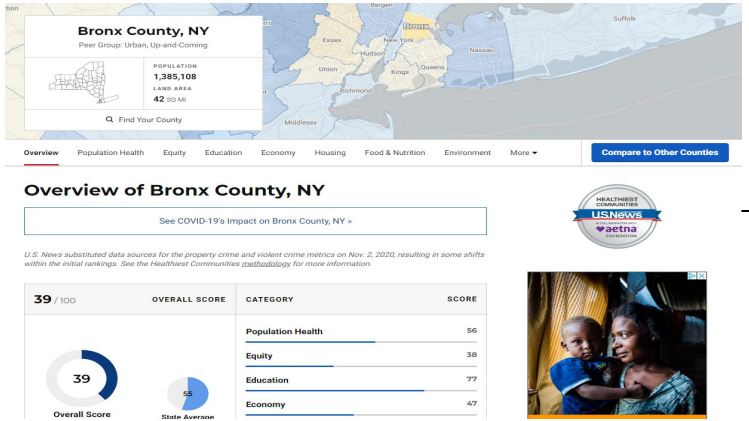
One of the biggest challenges with gentrification is that people are not aware that it is happening in a community until it is too late to address it and native residents have already been displaced. If local governments are aware that their communities are likely to become gentrified, they can allocate resources, gather community input and enact policies to avoid widespread displacement of native residents.

Solution

We developed a model to predict which U.S. counties are at risk of becoming gentrified in order to prevent the displacement of their native residents.

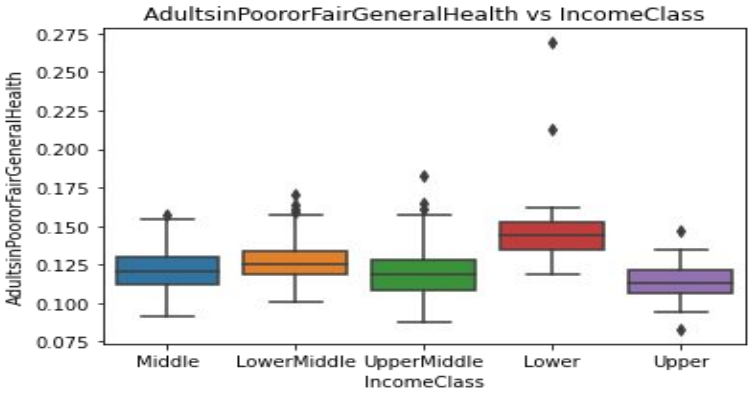
Dataset

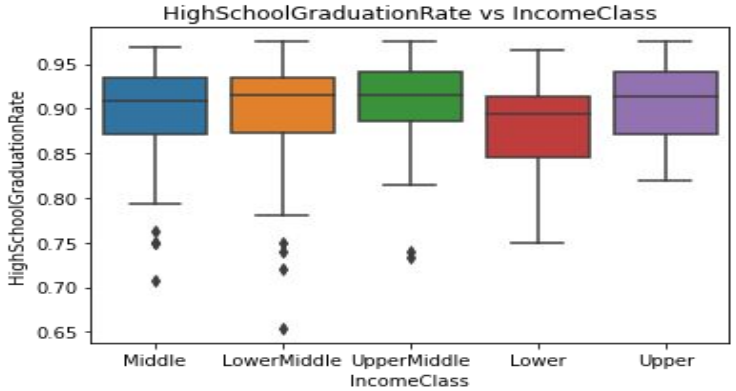
- 2020 Healthiest Communities Report
- Collaboration between the Aetna Foundation, U.S. News and the University of Missouri Extension Center for Applied Research and Engagement Systems
- Scores 3,000 counties on 84 indicators across 10 categories that drive overall community health
- Example variables for each county: **Adults in Poor or Fair General Health, Segregation Index Score, Gini Index, Affordable Housing Shortfall, Eviction Rate, etc**
- County profiles were web scraped using Python and converted to a county-level table
- Imputed all missing numerical values using KNN imputation



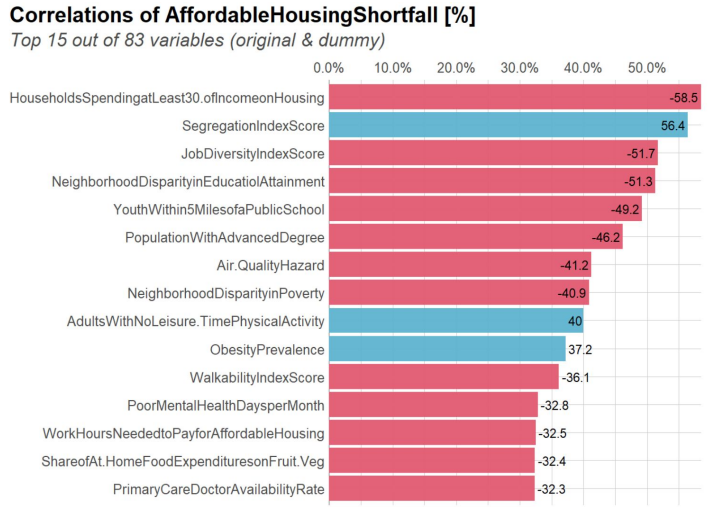
County_me	State_me	AccidentalDeathRate	AdultsinPoororFairGeneralHealth	AdultsWithNoLeisure-TimePhysicalActivity	AffordableHousingShortfall	Air-QualityHazard	AirToxicsExposure
ada-county	idaho	0.000378	0.1160	0.1460	-76.3	0.68	1.32
adair-county	iowa	0.000503	0.1170	0.2520	-28	0.24	
addison-county	vermont	0.000452	0.1140	0.1920	-60.5	0.25	4.1
alameda-county	california	0.000247	0.1270	0.1450	-66.2	0.53	1.44
albany-county	wyoming	0.000492	0.1600	0.1760	-72.9	0.22	2.13
albemarle-county	virginia	0.000326	0.1290	0.1690	-80.6	0.46	7.52
alexandria-city	virginia	0.000267	0.1310	0.1620	-67	0.56	0.8
allamakee-county	iowa	0.000563	0.1250	0.2130	-35	0.22	0.26
anne-arundel-county	maryland	0.000307	0.1120	0.2010	-57.1	0.68	4.94
anoka-county	minnesota	0.000377	0.1020	0.1920	-69.2	0.49	11.51
antelope-county	nebraska		0.1340	0.2330	-46.9	0.19	
arapahoe-county	colorado	0.000433	0.1200	0.1650	-79.2	0.51	4.61
archer-county	texas	0.000562	0.1320	0.1830	-39.5	0.35	1.18
archuleta-county	colorado	0.000489	0.1260	0.1220	-58.2	0.23	0.5
arlington-county	virginia	0.000217	0.1220	0.1390	-80	0.7	2.15

Exploratory Data Analysis





- We found disparities in General Health and Education Attainment between Income Classes
- The 'Affordable Housing Shortfall' variable had strong correlations with 'Segregation Index' and 'Households Spending at least 30% of their Income on Housing'



Model

- We used the **AffordableHousingShortfall** as our response variable/proxy for gentrification. It is defined as the "availability of affordable housing for families that earn less than 50% of median family income."
- Used 80% of our scraped data to train models and 20% to test them
- Trained 7 different linear and non-linear models: GLM, GLM Net, PLS, Neural Network, Basic Regression Tree, Random Forest, XGBoost
- Our GLM Net model was chosen as the final model because it had the lowest RMSE

Model	RMSE
GLMnet	17.60
PLS	18.11
Random Forest	18.82
GLM	19.20
XGBoot	20.01
Basic Regression Tree	24.05
Neural Network	31.55

Most influential predictors in our GLM Net model:

1. HouseholdsSpendingatLeast30.ofIncomeonHousing
2. SegregationIndexScore
3. JobDiversityIndexScore
4. PopulationWithAdvancedDegree
5. UnemploymentRate

We should pay attention to these factors in communities at risk of becoming gentrified

Conclusion

- Based on our work:
  - Community leaders in counties that the model identifies as at risk for gentrification (via high values of AffordableHousingShortfall) should take action
  - Programs should be established early on to create equitable ways for native residents to also benefit from any new resources in their neighborhoods.
  - Local economic initiatives can ease and prevent the displacement of residents before gentrification takes a strong hold of their community.

- In future iterations of our model, we should:
  - Web scrape all 3,006 U.S. counties rather than just a subset of counties for more training data (requires more computing power)
  - Web scrape data across different years of the report, not just 2020
  - Incorporate additional variables such as weather
  - Incorporate NLP from social media to flag what people may be saying about a community

Thank You!