Gentrification Prediction - Team 147

Ambasta, Harsh; Caballes, Ysabel; Long, U I; Mulla, Suhail; Vaddiparti, Tejasvi; Yao, Yang

Introduction

For most people, gentrification is an event that simply happens to them. Driven by economic forces (Redfern 2003), gentrifiers – primarily white and middle class – move into affordable neighborhoods, drive up prices, and eventually displace the culture and perhaps tenancy of long-term residents – who are primarily POC and lower income (Wyly and Hammel 2004).

This transformation has its pros and cons. On the one hand, gentrification can result in neighborhoods with greater racial diversity (Freeman 2009). But gentrification also results in higher income inequality (Freeman 2009), which has been linked to an “increase [in] both violent and property crime” (Atems 2020), as well as closures of long-standing businesses (Glaeser et al. 2023). Gentrification can also lead to negative health impacts, because of loss of services and common amenities (Anguelovski et al. 2021).

Aims and Objectives

Given the consequences of gentrification, our group aims to give more control to the people affected by giving them a map-based visual tool to predict a neighborhood’s gentrification levels, similar to that seen in Mubarak et al.’s paper (Mubarak et al. 2022).

By having access to this information, those living in areas on the verge of gentrification can start taking measures to preserve their neighborhood, and figure out ways to turn a potential disaster into an opportunity (Thurber 2021). On the other hand, those looking to move can more easily find areas with rich histories and soon to appreciate property values (Wilhelmsson et al. 2021), which would allow buyers to maximize their investment, and allow renters to weigh short-term affordability vs the long-term risk of rising rents (Aljohani 2023). This would also enable urban planners and governments to implement public policy to minimize the negative and heighten the positive effects of gentrification (Lees and Ley 2008).

In order to meet this objective, we plan to build an interactive map of different cities in the UK. The user would simply have to type in their postcode or zoom into their place in the map to find out whether their neighborhood is expected to gentrify or not. While that is our main objective, a side objective we would like to work on is to beat state of the art accuracy of other gentrification prediction models.

We are limiting our scope to the UK for three primary reasons: (1) quality census data, (2) extensive data on gentrification, being the birthplace of the term itself (Finio 2021), and (3) easily accessible house pricing and income data.

Ethical Considerations

While our goal is to help those affected by gentrification gain more information and control over their neighborhoods, we know that giving the wrong people access to this information could also exacerbate the phenomenon. In particular, corporate landlords are infamous for purchasing property in bulk in gentrifying neighborhoods and evicting tenants at a higher rate than small landlords (Raymond et al. 2016). Given their resources however, it is certain that corporate landlords are already using ML techniques to predict gentrification. Our paper will not give them any new information, but will make it available to individuals who did not have prior access.

There is also the risk of building inaccurate models that could result in ineffective or harmful public policy. To mitigate this risk, we will try a variety of approaches to find the most accurate one, and make sure to report any shortcomings in the model.

Literature Review

To predict gentrification, we first need to discuss how it is quantitatively measured. Classically, researchers have followed a two-step process: (1) identify if a neighborhood is eligible for gentrification, then (2) assess over time if the neighborhood gentrifies (Finio 2021). For example, using the Freeman methodology, a census tract is (1) marked eligible if (a) housing construction is below a metropolitan median, and (b) income is below the median, and (c) the tract is located in a central city. It is then (2) marked as gentrified if, after the measurement period, (a) there is a greater increase in educational attainment compared to the median and (b) an increase in real housing prices (Freeman 2005).

In recent years, researchers have also adopted ML methods to classify gentrification. PCA was used to combine measures that define gentrification, namely shares of people with higher educational attainment, higher occupational status, and house prices along with income to create a score that represents socio-economic status (Reade et al. 2019).

There is currently no standardized way to forecast gentrification. In a study comparing gentrification prediction models used by four US city governments, only 2 of 18 predictor variables were consistent for all models, and performance accuracies on a test city were highly varied, between 13.9% - 66% (Preis et al. 2021).

Models from academia are more promising. Stanford researchers trained a deep learning model on paired Google street view images to detect changes in infrastructure to predict gentrification with ~75% balanced accuracy (Huang et al. 2023). Deep learning methods are black boxes however and limit interpretability, which is important for public policy changes.

Researchers from Sydney (Thackway et al. 2023), Mexico City (Alejandro and Palafox 2019), and the UK (Gray et al. 2023) experimented with using tree based models to predict gentrification, achieving 74.7% balanced accuracy, 99.65% accuracy with 66% sensitivity, and 99.65% respectively. All three papers used census data, which is infrequent, to train the models whereas this group from Harvard used Yelp data and linear regression to show relevant correlation between certain businesses and indicators of gentrification (Glaeser 2018).

Research Design and Methods

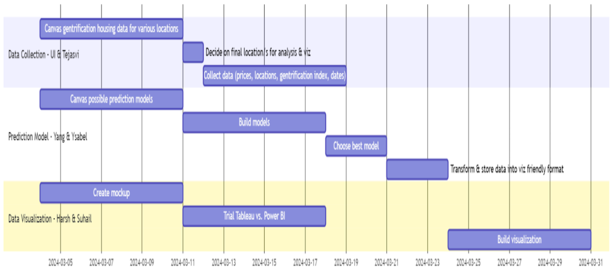


Figure 1: Previous Plan of Action

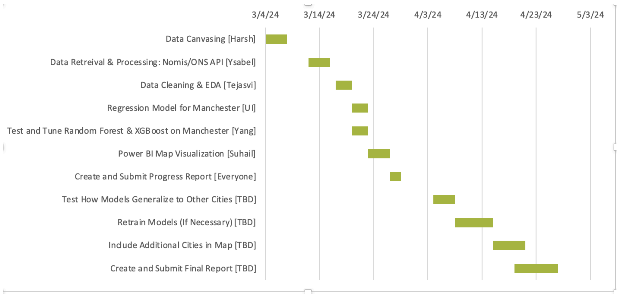


Figure 2: Revised Plan of Action

There was no ready-made dataset that we could use for our project. In order to build our dataset, we had to collect census data for 2001, 2011, and 2021 from Nomis, housing prices from the Office of National Statistics, geographic data from Geoportal UK, and income data from Gov UK. All of these sources are official web apps and APIs maintained by the UK government. We collected data at the lowest possible level of detail for each neighborhood, which is the LSOA. After collecting these tables, we merged them into one dataset and used OpenRefine to align variables between census years (each census has a slightly different set of questions and phrasing, resulting in slightly different tables for each year). We then scaled the data and used Reed's PCA method to create our binary gentrification prediction variable.

Once we had our dataset, we did some data exploration then tested out these boosting and bagging models: Random Forest, XGBoost, and LightGBM. We chose these because they are tree based models that result in lower variance, potentially lower bias for the boosted models, and they all inherently perform feature selection and weighing. The metrics we plan to use to measure performance are balanced accuracy, mean squared error (MSE), and mean absolute error (MAE). Balanced accuracy is necessary when there is an imbalance of data, while the latter two are useful additional metrics for comparing models.

For our visualization, we decided to use Power BI to build an interactive choropleth map. Our plan is to build the map so that each neighborhood is easily searchable, the coloring of an area denotes its level of gentrification, a tooltip appears with all relevant details when you hover over an area, and a slider allows the user to switch between different census years. We then plan to publish the map to the web and embed it into a Github page for easy access.

For our initial prototype, we decided to limit our dataset to just the Greater Manchester area instead of all UK cities. This allowed us to run our initial models while we were still collecting and preparing data for the rest of the UK, giving us a more thorough understanding of the whole process more quickly and efficiently.

Initial Results

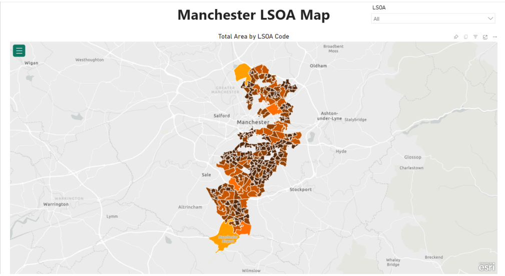
Before building our models, we wanted to get an idea on which variables are significant factors for predicting gentrification, so we performed an Ordinary Least Squares (OLS) regression on the dataset. Our aim was to determine variables with p-values less than 0.05. After testing various variable selections, the following variables were found to contribute the most to gentrification: Economic status, Ethnic group, Industry, Occupation, Tenure, Qualification, 2001 Density (number of people per hectare). The resulting model had an adjusted R-squared value of 0.72. This was done by categorizing LSOAs into gentrified or not comparing median house value to the Greater Manchester median house value.

We also observed that in gentrified areas, there is a higher proportion of people with no qualifications compared to non-gentrified areas. Conversely, we found that the presence of individuals with Level 4 qualifications and above is more prevalent in non-gentrified areas. These results highlight an intriguing pattern, and for our next step, we intend to conduct further research to better understand the underlying reasons behind these observations.

Shifting focus to the dependent variable, we observe that change in socio-economic score between 2001 and 2011 is normally distributed with standard deviation of 0.15. Using the 1 standard deviation threshold, there were 177 LSOAs that gentrified. Using the change in score as the variable to be predicted, we trained a Random Forest using python’s Sklearn library and XGBoost using the xgboost library. As a baseline, no hyperparameter tuning was done. The results for MSE, MAE, and balanced accuracy for RF were 0.027, 0.116, and 0.615. While XGBoost achieved 0.032, 0.131, and 0.582. Once the models were tuned, RF attained MSE, MAE, and balanced accuracy of 0.027, 0.114, and 0.651. Whereas XGBoost has similar MSE and MAE, but a balanced accuracy of only 0.5. For the next phase, we intend to further optimize XGBoost and to train LightGBM and a stacked model consisting of averaging predicted scores from RF, XGBoost, and LightGBM.

We still need to test our models to see how they perform in other large metro areas in England. This entails three additional tests: 1) testing performance on areas outside of Manchester without additional tuning and training 2) testing performance from 2011 to 2021 for Manchester and other metro areas 3) repeating (1) and (2) for newer models. Another important aspect we will be researching is how models trained on data in one city and year generalize to other cities, and how they compare when trained on the dataset of all cities in aggregate.

The initial map visualization has been built in Power BI. After finalizing the model to be used, the tooltip and actual map will be finalized. Below we have provided a map that has been developed using placeholder values. A hover over tooltip will be incorporated based on the results of the predictions from the model. In the top right corner, users have the ability to select LSOA by slicer.


Manchester LSOA Map

LSOA: All

Total Area by LSOA Code

Figure 3: Power BI Map of Manchester

**Everyone contributed equally to the project.*

Conclusions/Discussions

Under Construction

Works Cited

Alejandro, Yesenia, and Leon Palafox. “Gentrification Prediction Using Machine Learning.” *Advances in Soft Computing. MICAI 2019. Lecture Notes in Computer Science*, vol. 11835, 2019, https://doi.org/10.1007/978-3-030-33749-0_16.

Aljohani, Abeer. “Predictive Analytics and Machine Learning for Real-Time Supply Chain Risk Mitigation and Agility.” *Sustainability*, vol. 15, no. 20, 2023, https://www.mdpi.com/2071-1050/15/20/15088.

Anguelovski, Isabelle, et al. “Gentrification pathways and their health impacts on historically marginalized residents in Europe and North America: Global qualitative evidence from 14 cities.” *Health & Place*, vol. 72, 2021, https://doi.org/10.1016/j.healthplace.2021.102698.

Atems, Bebonchu. “Identifying the Dynamic Effects of Income Inequality on Crime.” *Oxford Bulletin of Economics and Statistics*, vol. 82, no. 4, 2020, pp. 751-782.

Finio, Nicholas. “Measurement and Definition of Gentrification in Urban Studies and Planning.” *Journal of Planning Literature*, vol. 37, no. 2, 2021, https://doi.org/10.1177/08854122211051603.

Freeman, Lance. “Displacement or Succession? Residential Mobility in Gentrifying Neighborhoods.” *Urban Affairs Review*, vol. 40, no. 4, 2005, pp. 463–91.

Freeman, Lance. “Neighbourhood Diversity, Metropolitan Segregation and Gentrification: What Are the Links in the US?” *Urban Studies*, vol. 46, no. 10, 2009, pp. 2019-2254.

Glaeser, Edward, et al. “Gentrification and retail churn: Theory and evidence.” *Regional Science and Urban Economics*, vol. 100, 2023, https://doi.org/10.1016/j.regsciurbeco.2023.103879.

Glaeser, Edward, et al. “Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change.” *AEA Papers and Proceedings,* 2018, https://doi.org/10.1257/pandp.20181034.

Gray, Jennie, et al. “Predicting Gentrification in England: A Data Primitive Approach.” *Urban Science*, vol. 7, 2023, https://doi.org/10.3390/urbansci7020064.

Huang, Tianyuan, et al. “Detecting Neighborhood Gentrification at Scale via Street-level Visual Data.” *arXiv*, 2023, https://arxiv.org/pdf/2301.01842.pdf.

Lees, Loretta, and David Ley. “Introduction to Special Issue on Gentrification and Public Policy.” *Urban Studies*, vol. 45, no. 12, 2008, pp. 2379-2384, https://www.jstor.org/stable/43197717.

Mubarak, Maryam, et al. “A Map-Based Recommendation System and House Price Prediction Model for Real Estate.” *ISPRS International Journal of Geo-Information*, vol. 11, no. 3, 2022, https://doi.org/10.3390/ijgi11030178.

Preis, Benjamin, et al. “Mapping gentrification and displacement pressure: An exploration of four distinct methodologies.” *Urban Studies*, vol. 58, no. 2, 2021, https://journals.sagepub.com/doi/10.1177/0042098020903011.

Raymond, Elora, et al. “Corporate Landlords, Institutional Investors, and Displacement: Eviction Rates in Singlefamily Rentals.” *FRB Atlanta Community and Economic Development Discussion Paper No. 2016-4*, 2016, https://ssrn.com/abstract=2893552.

Reades et al., 2019, “Understanding Urban Gentrification Through Machine Learning.” Urban Studies, vol. 56 Issue 5 pp. 922-942, 2019,https://doi.org/10.1177/0042098018789054.

Redfern, P.A. “What Makes Gentrification 'Gentrification'?” *Urban Studies*, vol. 40, no. 12, 2003, pp. 2343-2584.

Thackway, William, et al. “Building a predictive machine learning model of gentrification in Sydney.” *Cities*, vol. 134, 2023, https://doi.org/10.1016/j.cities.2023.104192.

Thurber, Amie. “Resisting gentrification: The theoretical and practice contributions of social work.” *Journal of Social Work*, vol. 21, no. 1, 2021, pp. 26-45. https://doi.org/10.1177/1468017319861500.

Wilhelmsson, Mats, et al. “Gentrification effects on housing prices in neighbouring areas.” *International Journal of Housing Markets and Analysis*, vol. 15, no. 4, 2021, pp. 910-929, https://www.emerald.com/insight/content/doi/10.1108/IJHMA-04-2021-0049/full/html.

Wyly, E., and D. Hammel. “Gentrification, segregation and discrimination in the American system.” *Environment and Planning A*, vol. 36, 2004, pp. 1215-1241.