Gentrification Prediction - Team 147 Final Report

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

Introduction

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).

Problem Definition

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), but expanded to all cities in the UK with census data available.

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.

Literature Survey

To predict gentrification, we first need to discuss how it is quantitatively measured. Currently there is 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).

Another method first put forth by Owens and adopted by Reades et al. is to perform

principal components analysis (PCA) on variables that encapsulate socio-economic change and take the first principal component to use as a socio-economic score (SES) for a neighborhood. Both Owens and Reades et al. used four variables specific to the UK census, proportion of residents with level 4 or higher qualifications, proportion in white collar jobs, median household income, and median house price (Owens 2012)(Reades et al. 2019). Similarly Thackway et al. used a similar approach except used the SEIFA score calculated by the Australian Bureau of Statistics which is taken from performing PCA on 21 variables that measure neighborhood socio-economic status (Thackway et al. 2023). Reades et al. and Thackway et al. discovered that one standard deviation (SD) of change in SES represented great enough change to be classified as gentrification.

As to how to predict gentrification, several studies in the past have made use of machine learning models with promising results. 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).

Methodology

In our research we did not come across any interactive map based tool utilizing machine learning to predict gentrification for all cities in the UK. Our main objective is to build such a map and our secondary objective is to outperform state of the art model's accuracy.

Dataset

To create our dataset we collected census data for all areas in the UK using the Nomis API, along with housing prices and income data from ONS surveys. Years when census data was collected were 2001, 2011, and 2021. Only the years 2011 and 2021 had all the necessary variables for our purposes and the lowest level of geography where we could include all variables was at the middle layer super output area (MSOA) level. On the data we performed exploratory data analysis (EDA) to ensure it is error free and to analyze some key variables. After understanding the data, we performed feature engineering by changing counts to proportions, normalizing variable names between the two years, and scaling the data using robust scaling.

Adopting the method employed by Reades et al. to classify gentrification, we performed PCA on proportion of residents with qualifications level 4 or higher, proportion in higher occupation, median household income, and median house price. Taking the difference of SES score (1st principal component) results in the dependent variable and any MSOA with one SD or more of change being classified as gentrified.

Initially, while we were still collecting and compiling data for all UK cities, we first limited our modeling dataset to just the Greater Manchester district and the years 2001 and 2011, so we could prototype our models and visualizations on smaller data. Once we had collected data for all cities, we retrained our models and built our final visualizations on the complete dataset using 2011 and 2021 values.

Modeling

Once the dataset was ready for modeling, we tested out Random Forest (RF), XGBoost (XGB), LightGBM (LGB), and a fully connected artificial neural net (ANN). These models are ubiquitous in data science as they have shown to perform well for a variety of prediction and classification tasks. RF is a bagging model that aggregates several decision trees and averages their output to decrease variance. XGB and LGB are boosting models that use decision trees as base models. However, they use trees built sequentially, where each subsequent tree improves on the error made by the previous tree. Both are able to reduce variance and potentially bias as well. The ANN is a deep learning model that uses several layers of affine transformations of features and non-linear activation functions to approximate a target function or probability distribution. Each model described performs inherent feature selection. One drawback is that each model requires tuning of various hyperparameters to optimize performance. Due to the continuous nature of the dependent variable, mean squared error (MSE) was the metric used for tuning and optimization. The metrics used to measure final performance were accuracy and balanced accuracy.

As part of the training process, the data was split into 30% testing and 70% training. Once testing was complete, we selected the best model and used the SHAP method to calculate the top five features that contributed most to the prediction outcome of each MSOA. SHAP relies on Shapley values and the calculation involves fixing a feature of interest, and calculating the dependent variable with various permutations of all features where features are included or excluded in each permutation.

Visualization

For the visualization, we used Microsoft Power BI. We created two maps based on (1) the results of gentrification for the year 2021 based on the data from the year 2011, and (2) the predicted gentrification for the year 2031 based on the data from the year 2021. The latter is our main visual.

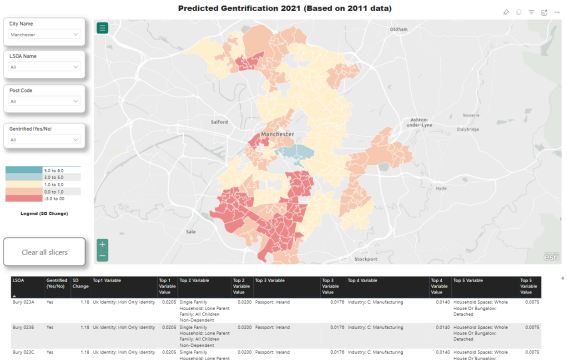

Figure 1: Map visual showing the results

In order for the visualization to render within Power BI, we needed to dimensionally model the data in order to get to the level of detail required. This involved using lookup tables to convert MSOAs to LSOAs and Postcodes, and imputing their gentrification values based on their corresponding MSOA.

The result is an interactive choropleth map that displays the geography and gentrification data at an LSOA level. Users have the option to use slicers to choose the city, LSOA, and Postcode, and then view the results of the modeling associated with that area.

Gentrification is visualized through color. The greater the gentrification, the more blue the neighborhood appears on the map. Conversely, neighborhoods that have/will become less gentrified are shown as red. In addition to this, there is a table visual below the map visual that allows the users to scroll through the available data for multiple areas.

Users can hover over each area and view more detailed information in a Tooltip. These details include whether or not that area was gentrified, the extent to which the area has been gentrified, and the top 5 variables responsible for these changes.

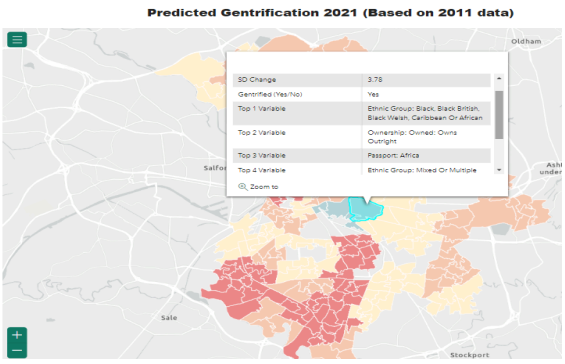

Figure 2: Tooltip on the map showing gentrification details for the area

Experiments & Evaluations

To start, we discuss some findings from the EDA concerning key attributes tied to gentrification for the top 3 wealthiest cities and the bottom 3 least wealthy cities ranked by median household income. More specifically, the attributes are qualification, occupation, industry, and ethnicity. Respectively, the top 3 and bottom 3 cities are London, Bristol, and Cardiff, and Coventry, Leicester, and Nottingham.

Qualification: Wealthier cities have a higher percentage of highly qualified individuals compared to the poorer cities. Among the top 3 cities, 37.5% and 19.5% of the population have Level 4 or above qualifications and no qualifications, respectively. In contrast, the bottom 3 cities have 22.7% and 27.6% with Level 4 or above qualifications and no qualifications, respectively.

Occupation: In the top 3 cities, the top three occupations are professional (20.5%), associate professionals and technical roles (15.9%), and managers and senior officials (14.0%). Conversely, in the bottom cities, the top three occupations are elementary (16.4%), professional (14.3%), and process, plant, and machine operatives (11.0%).

Industry: The top cities exhibit a more diversified distribution of industries, without dominance by specific sectors. Among the top 3 cities, the leading industries are wholesale and retail (13.6%), finance and insurance (10.8%), and professional and technical services (9.2%). On the other hand, the top three industries in the bottom cities are health and wholesale and retail (18.0%), health and social work (10.6%), and manufacturing (7.8%).

Ethnicity: In the top 3 cities, there is a higher proportion of Black, Black British, Black Welsh, Caribbean, or African individuals (11.4%) compared to the bottom 3 cities (5.8%). Conversely, for Asian, Asian British, or Asian Welsh individuals, with 15.4% in the top 3 cities compared to 20.0% in the bottom 3 cities.

Moving on to modeling. In the first experiment, predictions of which LSOAs would be gentrified in Manchester in 2011 were made using data from the 2001 census. Out of the 1,509 LSOAs, 177 LSOAs were actually gentrified using the one SD threshold. A RF and XGB model were tuned using grid search. Once the models were tuned, RF attained accuracy and balanced accuracy of 82.7% and 62.9%. Whereas XGB had similar accuracy, and balanced accuracy of 79.2%, and 53.5%.

Training on the full UK dataset of 212 features proved more fruitful. RF performed best relative to XGB, LGB, and ANN with balance accuracy of 79.3%. Which is better than what Reades et al. and Thackway et al. were able to get. Final model hyperparameters were 1,300 trees, max tree depth of 18, max features considered per tree as square root of 212, and min sample per leaf of 2. All model results can be viewed in Table 1. All accuracy measures were calculated by looking at gentrification of MSOAs between 2011 and 2021 as there has not been a census since 2021.

Model	**Accuracy**	**Balanced Accuracy**
RF	89.1%	79.3%
LGB	88.8%	78.2%
XGB	87.2%	76.1%
ANN	79.5%	59.2%

Table 1: Model Test Results

Since RF had superior performance, it was selected as our final model to make predictions for gentrified MSOAs in 2031 and for Shapley value calculations. From our predictions, in 2021 the top 3 cities with the largest share of gentrified MSOAs were London, Manchester, and Bristol. Roughly 77% of all gentrified MSOAs were in London, while Manchester and Bristol each had 5% of the gentrified MSOAs. Similarly, in 2031 it is predicted that 79% of the gentrified MSOAs will be from London with Manchester in second again with 4% of MSOAs and this time Birmingham in third with 3%. These predictions are not a huge surprise as gentrification is generally confined to large cities and the population of London is over 6x greater than the second most populous city in the UK.

From Shapley values we were able to understand what attributes of each MSOA contributed most to its prediction. Analyzing frequency of attributes in the most influential variable we discovered that in 2021, 54% of gentrified MSOAs had “Ethnic Group: Black” as the number one variable and 24% had “Single Family Household: Lone Parent Family.” Moving to 2031 predictions, “Ethnic Group: Black” was the most prevalent as the number one most influential attribute in 68% of predictions for gentrified MSOAs followed by “Passport: Africa.” This makes sense as a significant presence of minority groups in a neighborhood are part of the preconditions for gentrification to happen. Additionally, for 2031 in Leeds, Luton, and Salford there will be a significant percentage of gentrified MSOAs.

Conclusions & Discussions

In summary, we developed an interactive choropleth map that predicts gentrification of MSOAs in the UK for the years 2021 and 2031 using machine learning models trained on census data. The machine learning algorithm underlying the predictions is a

random forest which we were able to tune to 89% accuracy and 79% balanced accuracy on test data. In addition to gentrification predictions, the map features a tooltip on hover with Shapley values per LSOA that informs the viewer of which census attribute most contributed to the prediction. In that way users can understand the reason a neighborhood may become gentrified.

We initially wanted to use the "Publish to Web" feature in Microsoft Power BI to publish the report as an interactive web page that could be used to choose the post code and view the gentrification results. We were able to create the report but the free version of Power BI does not satisfy some of the requirements needed to render the visual. Additionally, even with a paid account, every web visual can only contain a maximum of 30,000 locations. We were therefore unable to publish to web. But we have included the Power BI report (*.pbix file) in our submission, which is easily accessible with the free Power BI Desktop application.

Team Members Efforts

Harsh: 1) edit proposal, 2) canvas data sources

Ysabel: 1) draft & edit proposal, 2) collect & clean final dataset, 3) feature engineering, 4) edit progress report, 5) edit final poster, 6) edit final report

UI: 1) edit proposal, 2) edit proposal presentation, 3) initial data collection, 4) EDA for initial and final datasets, 5) tuned XGBoost, 6) edit final poster

Suhail: 1) edit proposal, 2) canvas & plan visualization method, 3) edit progress report, 4) create map visual, 5) post-map EDA, 6) edit final report

Tejasvi: 1) edit proposal, 2) draft & edit proposal presentation, 3) collect & clean initial Manchester dataset, 4) LightGBM modeling, 5) draft & edit final poster, 6) post-map EDA

Yang: 1) edit proposal, 2) edit proposal presentation, 3) create proposal video, 4) XGBoost, RF, & ANN modeling, 5) draft progress report, 6) draft & edit final report

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.

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.

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.

Owens, Ann. “Neighborhoods on the Rise: A Typology of Neighborhoods Experiencing Socioeconomic Ascent.” *City and Community,* vol. 11, issue 4, 2012, https://doi.org/10.1111/j.1540-6040.2012.01412.x.

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.

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 neighboring 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.