Developing an Early-Warning Gentrification System with Machine Learning

1053033

August 16, 2021

Thesis submitted in partial fulfilment of the requirement for the degree of MSc in Social Data Science at the Oxford Internet Institute at the University of Oxford

Word Count: 12,238

Contents

Title: Developing an Early-Warning Gentrification System with Machine Learning
Author: 1053033
Word Count: 12,238

Abstract

First coined to describe the process of "invasion" of working-class neighbourhoods in London by the middle class, gentrification has been thoroughly studied in the social sciences ever since. This phenomenon involves two essential parts: a 'class-based colonization' of inexpensive neighbourhoods and reinvestment in housing stock. While scholars disagree about whether displacement represents a fundamental part of gentrification, it is certainly a potential – and important – consequence. Residents may have to choose between the negative impacts of displacement or spending an increasingly higher share of their income on housing; understanding the dynamics of gentrification and which neighbourhoods are likely to go through the process may aid governments in mitigating the negative impacts of the process. However, most traditional measures of gentrification are collected too slowly to be helpful to policymakers before it is too late. This study aimed to build an 'early-warning gentrification system' using data from Twitter and other sources accessible to policymakers, including 311 complaints, crime rates, and address vacancies or changes, in addition to socioeconomic features commonly used in past attempts to forecast gentrification, to predict which census tracts would gentrify from a set of 405 eligible tracts in NYC between 2010-2018. From Twitter, I extracted both structured and unstructured features. Using a Mallet LDA topic modeller, I estimated the probability that the tweets in a given month-year-census tract matched the topics learned by the model. I then built random forest and logistic regression models that were tuned using tenfold cross-validation to predict gentrification. The full model performed the best, greatly improving on the model with non-Twitter features, the baseline model, and random chance. An investigation of the importance of the features in the model also revealed links between gentrification and certain topics, such as employment, politics, sports, or the use of slang words.

1 Introduction and Literature Review

1.1 Gentrification: Background

Gentrification, a term first coined by sociologist Ruth Glass (1964) to illustrate the relatively new phenomenon of middle-class Londoners moving into and displacing the original residents of working-class neighbourhoods, has inspired much debate in the social sciences.

Nebulous by nature, gentrification often elicits the famous adage, 'you know it when you see it.' No consensus definition exists, with researchers' definitions often differing depending on their field of work and research question. Nevertheless, most research tends to converge on a two-part definition: reinvestment in derelict or working-class residential buildings as well as a 'class-based colonization' of affordable neighbourhoods – in other words, middle- or upper-class individuals moving into neighbourhoods previously inhabited mainly by the working class.

Some researchers focus on capital and the built environment when defining gentrification, with one defining it as '[occurring when] communities experience an influx of capital and concomitant goods and services in locales where those resources were previously non-existent or denied' (Prince, 2014). To measure gentrification, many studies have relied on average housing value or rents, with large neighbourhood increases often heralding the onset of the process (Jain et al., 2021; Preis et al., 2021; Edlund et al., 2016; Reades et al., 2018).

Others emphasize the second half of the definition, arguing that the identity of new in-movers into a neighbourhood is the driving factor behind gentrification. Such researchers have used various proxy measures as indicators of gentrification in action: the average income of a neighbourhood is a common one (Jain et al., 2021; Chapple and Zuk, 2016; Zuk et al., 2017; Reades et al., 2019). So too is the percentage of residents with a bachelor's degree (Glaeser et al, 2018; Jain et al, 2021; Zuk et al., 2017; Reades et al, 2018) or the percentage of residents between 25 and 34 (Glaeser et al, 2018), in recognition of the fact that many young but highly educated individuals, who often represent the 'first wave' of gentrifiers, are just beginning their careers and do not yet earn a high income. Race is also commonly used in studies on gentrification in America, due to its close association with income and class – an increase in the white population in Hispanic or black neighbourhoods often serves as a warning for gentrification (Preis et al., 2021). Other, less common variables used as proxies of gentrification include the share of non-English speakers or households that rent, the number of rent-burdened households, average household size, or type of occupation (Preis et al., 2021; Reades et al., 2018; Clark, 2005; Davidson and Lees, 2005; Ding et al., 2016; Freeman, 2005).

One strand of thought points to the role of the 'creative class'– professionals who create 'meaningful new forms', in creative industries as well as traditional ones like finance, technology, healthcare, and business management – in economic prosperity, and consequently, gentrification; these young, upwardly mobile professionals often move into working-class, affordable neighbourhoods, causing an increase in rent and attracting amenities that may fundamentally change the look and feel of a neighbourhood (Florida, 2002; Wainwright, 2017; Rotondaro, 2019; Zuk et al., 2017).

1.2 Consequences of Gentrification

While researchers disagree about whether gentrification always leads to displacement – when households must move due to conditions beyond their control that are impacting their building or neighbourhood (such as increased rent prices), it is certainly a potential – and important – consequence (Atkinson, 2000a; Atkinson, 2003). Some researchers and anti-gentrification activists argue for a direct causal link between gentrification and displacement; for example, Atkinson (2000a) combines cross-sectional and longitudinal census data in London to demonstrate a loss in low-income and elderly residents in formerly working-class areas after an increase in professionals in those neighbourhoods, arguing the latter caused the direct displacement of the former groups. Anecdotal reports of landlords intentionally forcing out low-income tenants to re-rent their rooms or sell their places for redevelopment intended to attract higher-income tenants are common (Chong, 2015).

Other research is much more sceptical of the alleged direct displacement effect of gentrification. Using longitudinal Medicaid records to trace the moves of low-income children in New York City over a seven-year period, Dragan et al. (2020) find that gentrification is not associated with a significant increase in household moves; instead, they find that low-income families move frequently, regardless of whether their neighbourhood is gentrifying. Moreover, families that stay in gentrifying neighbourhoods experience larger enhancements to their residential environment than those in neighbourhoods that start out – and stay – low-income. However, those that move out of a gentrifying neighbourhood tend to move further away, indicating they might need to travel longer distances to find affordable housing.

Regardless of whether it directly displaces low-income households, gentrification clearly leads to radical neighbourhood changes that can alienate current residents. Even if low-income households are not directly displaced by gentrification, they can be supplanted by higher-income ones in the medium- to long term (Atkinson, 2003; Zuk et al., 2017); as low-income households break up or move for non-gentrification related reasons (e.g., downsizing by families whose children leave their household, upsizing by families looking to add members, flatmates looking to have their own place, etc.), they are replaced by higher-income ones. Over time, this can lead to radical changes in the households that compose most of the neighbourhood. This loss of community can be devastating for households, especially those comprised of elderly individuals, many of whom are more negatively affected by social changes around them (Atkinson, 2000b). Similarly, the amenities and businesses that characterise these gentrifying neighbourhoods also change to suit the needs of their new residents, leading to the closure of old neighbourhood favourites and excluding long-term residents (Chong, 2017; Lung-Aman, 2021; I. Gould Ellen, personal communication, 15 April 2021).

1.3 Now- and Forecasting Gentrification

Since most measures of gentrification rely on census data updated once a decade (or five-year estimates updated every five years), official recognition be much slower than the actual process of gentrification. Policymakers and programs that minimize the negative impacts of gentrification could thus greatly benefit from more frequently updated gentrification measures.

One of the first attempts at a neighbourhood-level ‘early warning’ system was created in 1984, by the Center for Neighbourhood Technology in Chicago, which constructed a portal of property data such as code violations and tax delinquencies intended to monitor the housing conditions at a neighbourhood level, allowing intervention by policymakers if necessary. Ever since, cities and academics have attempted to create systems that can predict a neighbourhood’s susceptibility to gentrification or ‘now-cast’ neighbourhood change in real time using socioeconomic and environmental factors.

Common variables used include crime rates, number of 311 calls, number of parks, walkability, accessibility to public transport, household racial or economic composition, and so on (Chapple and Zuk, 2016; Preis et al., 2021). These models usually rely on techniques ranging from simple correlation to multivariable linear regressions. Unfortunately, these systems usually produce inconsistent and/or inaccurate outcomes; for instance, when researchers applied the methods of four different cities' gentrification-warning systems to Boston, they produced four different maps of gentrification-related displacement risk (Chapple and Zuk, 2016)! Similarly, are thus unable to be used by policymakers to target resources or make decisions. More recent work has turned to machine learning techniques like random forest models to predict gentrification from fixed census variables, greatly improving the accuracy from the traditional models (Reades et al., 2018; Gardiner and Dong, 2021).

Following in the footsteps of work on urban segregation, mobility, and social diversity showing the utility of geotagged social media sites like Twitter, Foursquare, and Instagram (Cagney et al., 2020; Cranshaw et al., 2012; Boy & Uitermark, 2016), a few studies have attempted to 'nowcast' gentrification – or provide a real-time assessment of whether a neighbourhood is gentrifying – using social media data. Such data can provide real-time information about the activities and movements of individuals, unlike many traditional data sources, and they are far more easily collected than traditional sources. Glaeser et al. (2016) use Yelp reviews to show that the number of businesses listed on Yelp is strongly associated with gentrification, which they define as an increase in the share of the population that is white, university-educated, or between the ages of 25 and 34. Similarly, Jain et al. (2021) scrape Airbnb reviews and demonstrate that several features, both structured ones like the number of listings or reviews in a neighbourhood, the number of bedrooms and price of said listings, and overall ratings, as well as unstructured ones like the topics or sentiment present in the reviews, are associated with their aggregate measure of gentrification. Poorthuis et al.'s study (2021) is most similar to my work; they use Twitter data to examine the mobility patterns of residents of Lexington, Kentucky, in an attempt to shed light on gentrification, finding that visitors to gentrifying neighbourhoods are whiter and more highly educated than the current residents.

1.4 Significance and Overview of Work

This work combines not only social media data and the socio-economic features used in many previous studies on predicting gentrification, it also uses several non-socioeconomic behavioural sources that have individually been shown to impact gentrification. Moreover, it is the first (to my knowledge) to use a topic modeller to investigate the potential use of Twitter discussion in gentrification. Although Jain et al. also used topic modelling, Twitter discussions are far more broad and not restricted to renting short-term accomodation, so they may be able to provide more insight into gentrification. Moreover, although Poorthuis et al. also used Twitter data, they examined mobility of users rather than the language that users used. Unlike most studies on gentrification, it uses only data that could feasibly be used for free by policymakers. As such, this study contributes meaningfully to the literature.

Section 2 describes the data, including both its collection and pre-processing performed on it. Section 3 discusses the methodology of the work, including how models were chosen. Section 4 describes the results of the model performance, comparing the full Twitter model to the baseline and non-Twitter feature model. Section 5 discusses the results, putting them in context of the current literature. Finally, Section 6 concludes, summarising the work and discussing its limitations and potential future research.

2 Data Collection and Pre-Processing

2.1 American Community Survey: Socioeconomic Variables & Gentrification

To extract socioeconomic features and determine which census tracts were gentrified, I used census-tract level data from the U.S. Census Bureau's American Community Survey (ACS), which surveys a random selection of approximately 3.5 million households annually to monitor demographic changes more quickly than the decennial census. The use of the ACS's five-year estimates is necessary to obtain accurate census-tract level data. I downloaded census-tract level data on educational attainment, race and ethnicity counts, median household income, gross rent, and owner-occupied house values, household income range counts, and number of households renting from the 2006-2010 and 2014-2018 ACS five-year estimates using IPUMS, a population database that includes microdata samples from the U.S. Census and ACS (Manson et al., 2020).

The Census Bureau changes its tracts slightly every ten years to account for population change; both the 2006-2010 and 2014-2018 estimates (hereafter referred to as /textit2010 ACS Estimates and /textit2018 ACS Estimates) were aggregated to census level using the 2010-2020 census tracts. This results in 2167 census tracts for New York City, which comprise the five boroughs of Brooklyn (Kings County), Manhattan (New York County), Queens (Queens County), Staten Island (Richmond County), and the

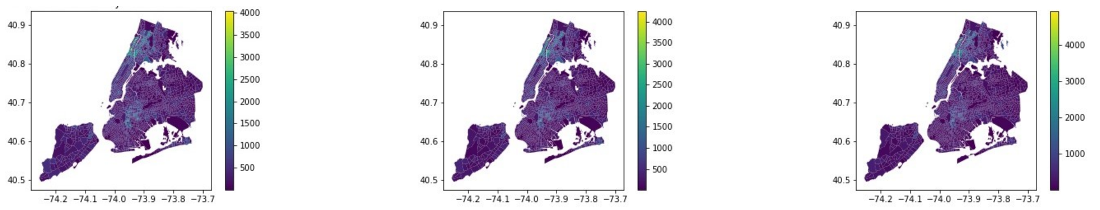
Bronx (Bronx County). I dropped all census tracts with zero population (53), as well as those missing estimates necessary to determine the gentrification labels (see below), resulting in 2098 census tracts. I also adjusted all nominal dollar estimates to 2018 dollars.

2.2 Twitter

Twitter allows its users to geotag their exact geographic location (latitude and longitude up to six decimal points, which corresponds to a point about 10 cm2), although Twitter removed the ability to precisely geotag in 2019 as geotagging frequency had fallen dramatically and it raised privacy concerns (Porter, 2019). I relied solely on Twitter data collected between 2011 and 2014, so geotagged tweets were still frequent.

Using Twitter's API v2 for Academics (Academic Research Product Tract, 2021), which allows researchers to collect up to 10 million tweets per month for free, I collected every tweet from every Wednesday and Saturday in 2011-2013 that was geotagged within five bounding boxes approximately corresponding to the whole of NYC. Wednesday and Saturday were chosen to reflect average weekday and weekend discussion, respectively; I did not use every day of the week as it would result in too much data to reasonably analyse. The bounding boxes also included tweets tagged from outside NYC, which were later deleted. I downloaded tweets from 2011-2013 to avoid any overlap with the labelling of neighbourhoods that were eligible to be gentrified (from the 2010 ACS Estimates) and those that did gentrify (from the 2018 ACS Estimates, which was collected between 2014-2018). The dataset consists of 5.79 million tweets in total, with approximately 4.2 million of those labelled as English-language tweets.

Although any dataset collected from Twitter suffers from potential selection bias – Twitter users are significantly younger, more highly educated, higher income and more politically left wing than the US population as a whole (Wojcik & Hughes, 2019) – this should not damage the predictive power of the gentrification model. The very characteristics that distinguish Twitter users from the general public are also closely associated with gentrifiers – educated, young, white, creative (Florida, 2002; Glaeser et al., 2018; Edlund et al., 2016); Twitter may thus serve as an effective early warning of gentrification.



(a) 2011 (b) 2012 (c) 2013

Figure 1: 311 Counts by Year

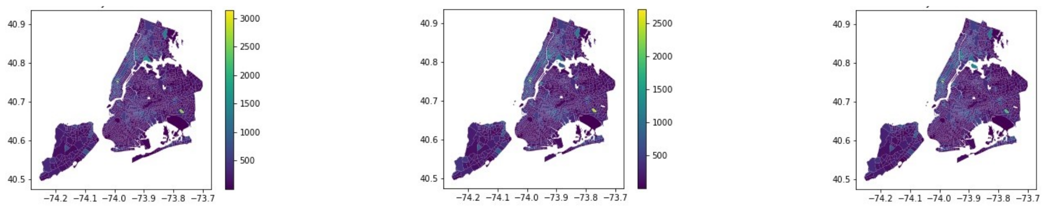
2.3 311 Calls

311 handles all requests for non-emergency public services from NYC residents; calling the number allows residents to access NYC services and put in complaints to the relevant authorities. Importantly, residents frequently use 311 to report noise or other quality-of-life complaints, which are frequently responded to by the police (NYC 311). Past research has shown that residents in racially diverse neighbourhoods – often an indicator of a recently gentrified neighbourhood – are more likely to call 311 than other areas (Legewie & Schaeffer, 2016). Moreover, one analysis of recently gentrified NYC neighbourhoods found a remarkable increase in quality-of-life 311 calls (usually noise complaints) after the onset of gentrification (Thuy Vo, 2018).

Using the Socrata Open Data API, I downloaded data on 311 complaint type, complaint location (latitude and longitude), and date from NYC's 311 database ('311 Service Requests from 2010 to Present') for 2011-2013. I downloaded the following complaint types due to their possible association with gentrification: noise; bike rack conditions, as wealthier households, which are likely to drive gentrification, bike more ('Cycling in New York City, 2007 to 2014', 2016), and for hire vehicle complaints, as Uber (and most likely other for-hire vehicle companies as well) serves more individuals in wealthy or gentrifying neighbourhoods (Rubenstein, D., & Cheney, B., 2015). Figure 1 shows the total count of 311 calls by census tract in each year. Duplicates and any that did not contain values for latitude/longitude were dropped.

2.4 Crime Rates

Previous research has demonstrated that falling crime rates may cause gentrification, as higher-income and more educated individuals feel safer moving into – and investing in –neighbourhoods previously characterised by high crime and disinvestment (Gould Ellen et al., 2019). I downloaded crime data between 2007-2013 on type of crime,



(a) 2011 (b) 2012 (c) 2013

Figure 2: Crime Counts by Year

whether the crime was a felony, and crime location (latitude/longitude) from NYC's Open Data database using the Socrata Open Data API. Any reports that contained null values were dropped. Figure 2 shows crime rates by census tract.

2.5 Changes in Addresses & Vacancy Rates

Given that gentrification is characterised by reinvestment in its housing stock – and possibly increased turnover in commercial businesses to meet the differential needs of the new residents of the neighbourhood – frequent changes in addresses might indicate gentrification. To measure this, I used the Department of Housing and Urban Development (HUD) Vacant Addresses dataset, which is aggregated quarterly at a census-tract level by the United States Postal Service and can be provided to governmental and non-profit organisations (including researchers at universities or in the government). Frequent residential and commercial turnover should be reflected by increased short-term vacancies, while increased construction should be reflected by a higher number of addresses.

This dataset contains information on total count of residential and commercial vacant addresses, average number of days vacant, and number of residential or commercial addresses vacant for 0-3 months, 3-6 months, 6-12 months, 12-24 months, 24-36 months, and 36 months or more. I used data from 2012 and 2013, as all prior datasets were tabulated using 2000 census tracts and cannot be estimated for the 2010 census tracts.

2.6 Census Tracts and R-Trees

The crime, 311, and Twitter data were geotagged using latitude and longitude; however, the data need to be associated with a census tract to properly extract census-tract-level features. Linking each individual tweet, crime, or 311 report to its corresponding census tract is too computationally intense to accomplish in a reasonable timeframe; instead of examining each individual tweet/report, I instead found all the tweets and/or reports corresponding to each individual census tract with the R-trees algorithm, which radically accelerates geographical indexing. R-trees group together nearby points or geographical objects (in this case, the latitude/longitude coordinates of each tweet/report) to form a 'minimum bounding box.' Nearby minimum bounding boxes are then grouped together to form larger bounding boxes, and so on, until only one top-level bounding box remains. To search, the algorithm inputs a query box (in this case, the census tract) and finds the top-level bounding boxes that intersect the tract. The algorithm then finds the 'child' bounding boxes inside each matched top-level bounding box that intersect with the census tract, then applies the same process to the matching 'child' bounding boxes, and so on until it reaches the lowest level (in this case, the latitude/longitude points of the individual tweets or reports). By applying the R-trees algorithm to each census tract in New York City, I could match every tweet/report to its corresponding tract in a matter of hours rather than days.

3 Methodology

3.1 Determining Gentrification Labels

As previously discussed, multiple definitions of gentrification exist. Drawing on the work in one of the Chapple references and the Urban Displacement Project, I used the following definition: A census tract in the 2010 ACS Estimates is eligible to gentrify if:

- Its gross real median rent or house value is < 80% of the NYC median And 3 of the following 4 apply:
- The share of households that are low-income (earning under $45k in 2010 dollars)[1] > NYC median
- The share of residents 25 or older with at least a bachelor's degree < NYC median
- The share of households that rent > NYC median

[1] 'Low-income' is defined (following Chapple et al., YEAR) as households making under 80% of NYC metro area median income for a single person (approximately the 40th percentile in household income), which was $44,350 in 2010. I rounded this to 45,000 to match the ACS estimates ranges.

- The share of residents that are nonwhite (including white Hispanic or Latinos) > NYC median

A census tract in the 2018 ACS Estimates has gentrified if:

- It is eligible to gentrify in 2010
- Its change in university-educated residents > NYC median change
- Its change in real median household income > NYC median change
- Either change in median real rent or in median value for owner-occupied homes > NYC median change

Of the 2098 census tracts considered, 405 (19%) are eligible to gentrify, 97 of which (24%) did so. Figure 3 shows the distribution of eligibility and gentrification in NYC.

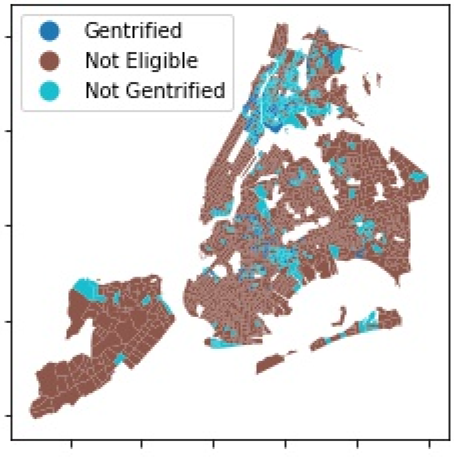


Figure 3: *Map of Gentrification by Category in NYC*

3.2 Feature Extraction

3.2.1 Socioeconomic Features

As a baseline model, I used a predictive gentrification model with only socioeconomic features. In addition to the socioeconomic variables necessary to define gentrification, which included percent of the population with at least a bachelor's degree, percent non-white (including all Hispanic and Latinos, regardless of their race), percent of households renting, median rent, median home value, percent of households considered low-income, and median household income, I also calculated the percent of the population that is black, white, or unemployed, for each census tract. Guided by both previous research (Gardiner & Dong, 2021; Jain et al, 2021; Glaeser et al., 2018; Chapple & Zuk, 2016) as well as a correlation matrix indicating the correlations between the possible socioeconomic features for all census tracts eligible to gentrify (as demonstrated by the heatmap in Figure 4), I chose percent of the population that is black, white, unemployed, and with at least a bachelor's degree, in addition to median rent and percentage of households that rent in a given census tract.

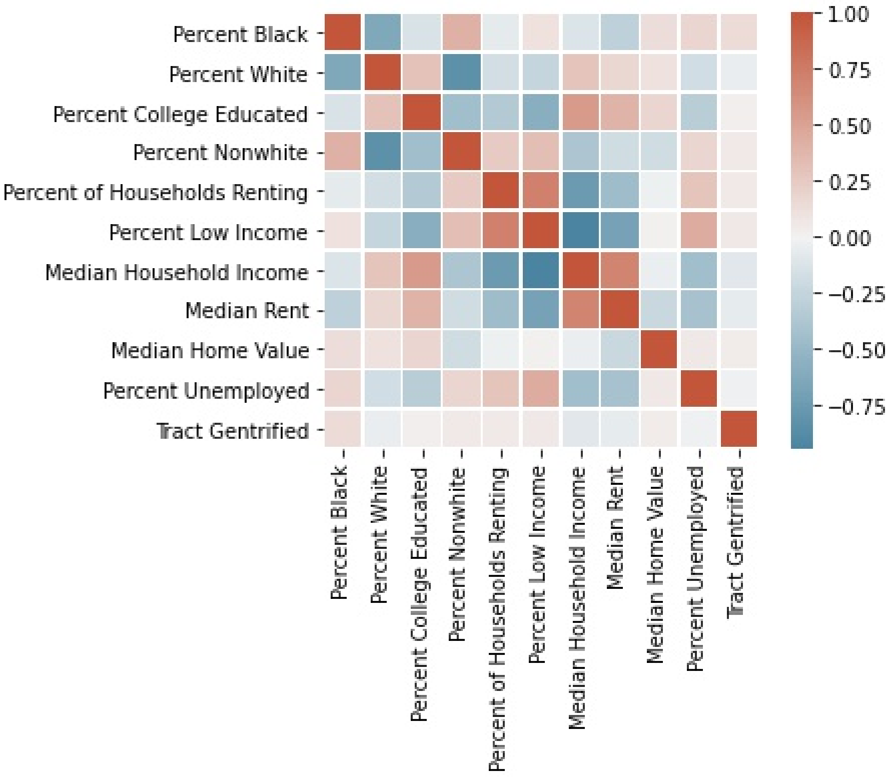


Figure 4: *Socioeconomic Features Correlation Heatmap*

3.2.2 Non-Twitter Features

311 Complaints

To measure whether noise, for hire vehicle, and bike rack condition complaints increased between the two time periods used to determine whether gentrification occurred (2006-2010 and 2014-2018), rather than use individual time period counts, which could be subject to random variance, I instead measured the 2011-2013 trend in monthly per-capita complaints for each type of complaint in each census tract in the model. To do so, I first collapsed per-capita count into month-year-census tracts, then replaced any missing months with 0, resulting in 36 count observations (3 years * 12 months in each year) for each census tract. I then calculated the line of best fit using the following linear regression:

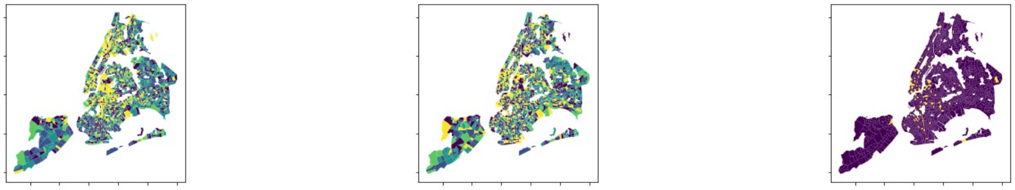
$$Com_i = B_i * time_it \tag{1}$$

where Com_i equals the complaint per capita count, $_i$ is the trend line for census tract i, and $time_it$ is the number of months since January 2006 for census tract i. Regressing Com_i on $time_it$ thus provided a trend line for each individual census tract that represented the increase (or decrease) in per capita complaints over the time period. Each tract's $_i$ for all three complaint types – noise, for hire vehicle, and bike rack condition – were included as features in the model.

The maps in Figure 5 show per-capita trends in each type of complaint between 2011-2013, with lighter colours representing higher trends in per-capita complaints. Noise and for-hire vehicle complaint trends are mapped by quantile, while bike rack trends is mapped by positive and negative trends (with purple representing negative trends and yellow positive trends). Yellow and green demonstrate positive growth in complaints, while purple, dark blue, and light blue represent negative trends. As the maps indicate, noise complaints were more likely to grow than other types of complaints; in addition, growth in noise and for-hire vehicle (to a lesser extent) complaints seems to be broadly reflective of gentrification trends.

Crime Rates

I similarly calculated the per-capita crime rate trends for each individual census tract. However, while 311 complaints most likely only reflect a change in the current composition of a neighbourhood and thus represent a sign of gentrification in progress, a change in crime rates could be both a harbinger of gentrification to come and a sign of its current occurrence. The young, educated, upwardly mobile residents that jump-



(a) Noise (b) For-Hire Vehicles (c) Bike Rack Condition

Figure 5: Complaints by Census Tract

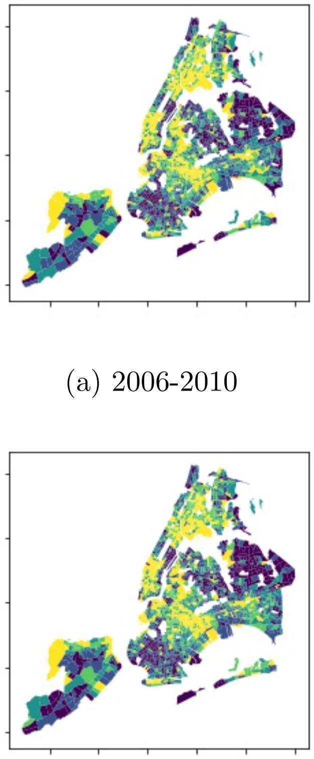
start the process of gentrification may be attracted to neighbourhoods with a recent decrease in crime, while a neighbourhood with a recent influx of such residents – in other words, one that is currently gentrifying – may see a further decrease in crime per capita, particularly more serious crimes like felonies, due to the increased population and changed composition of the neighbourhood. I thus included crime rate trends from 2006-2010 – the years immediately preceding the gentrification period – as well as from 2011-2013. To capture the potential additional impact on neighbourhood perceptions of particularly serious crimes, I also included felony trends between 2006-2010 as well as 2011-2013.

In addition to per-capita crime rates, absolute count of crime may also contribute to the perception of a neighbourhood's safety. Some neighbourhoods demonstrate high per-capita crime despite being seen as 'safe' places due to the relatively low number of residents and high number of tourists (e.g., Midtown in NYC). Thus, trends in felony counts and total crime counts between 2006-2010 as well as 2011-2013 were also included in the model.

Figures 6 and 7 demonstrate trends in crime per capita and total crime between 2006-2010 and 2011-2013. As with 311 complaints, crime rates are separated into quantiles and lighter colours indicate greater trends. The maps demonstrate that crime per capita increased in much of Brooklyn between 2006 and 2010 (possibly due to the negative impacts of the 2008-2009 Great Recession), continuing to increase in many (though not all) of these tracts between 2011 and 2013; however, total crime count dropped.

Change in Addresses Features

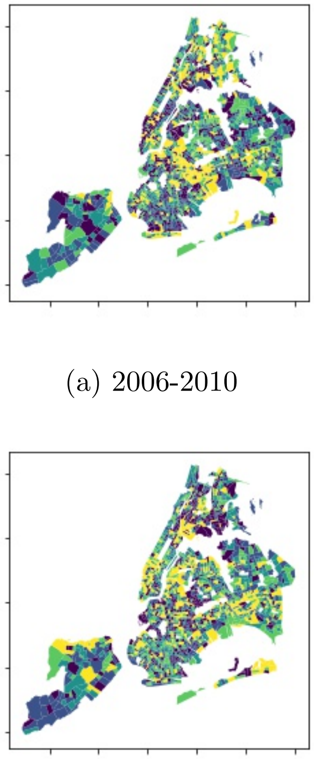
To capture potential indicators of gentrification like reinvestment in housing stock, high turnover in commercial enterprises, and short- or long-term vacancies in residen-



(a) 2006-2010

(b) 2011-2013

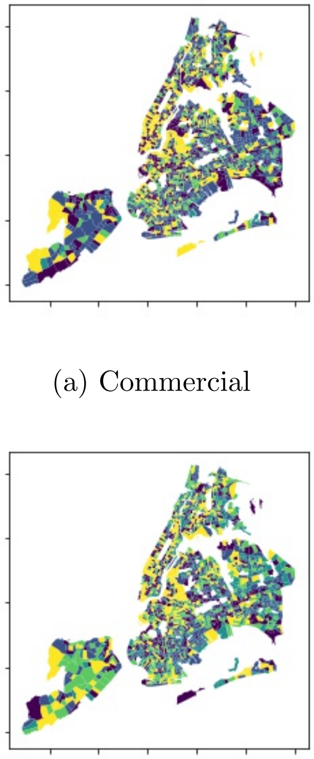
Figure 6: Crime per-Capita Trend



(a) 2006-2010

(b) 2011-2013

Figure 7: Crime Count Trend



(a) Commercial

(b) Residential

Figure 8: Addresses Count Trend: 2012-2013

tial or commercial buildings, the full model contained features extracted from the HUD dataset, including total number of commercial addresses, total residential addresses, total number of short-term (defined here as vacancies shorter than six months) residential as well as commercial vacancies, and total number of long-term (vacancies longer than six months) residential as well as commercial vacancies. As with the 311 and crime rate features, I calculated the census-tract level trend line for each variable, between 2011 and 2012 (due to data availability), for use as a feature in the model. Figure 8 shows the growth rates for the total number of commercial and residential addresses in New York.

Non-Twitter Feature Correlations

Figure 9 shows the correlation between the non-Twitter features and gentrification label for census tracts eligible to gentrify; a brief glance reveals that noise reports, crime, and residential addresses are most strongly correlated to gentrification, while vacancies are negatively related.

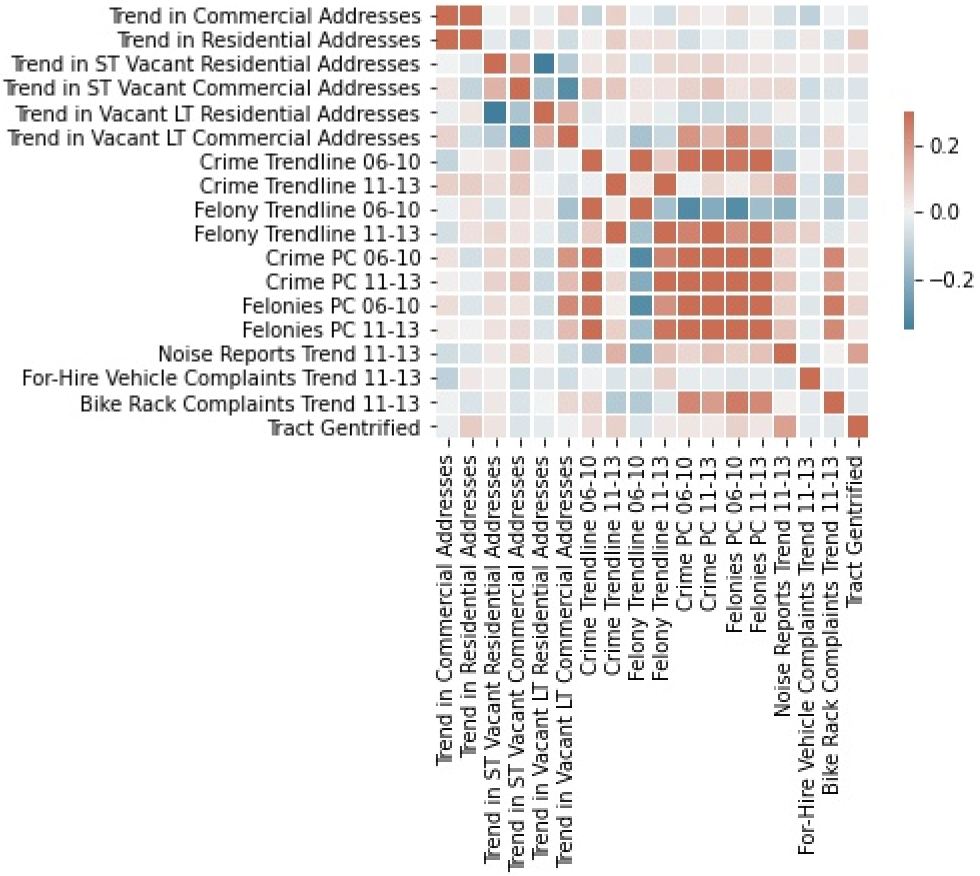
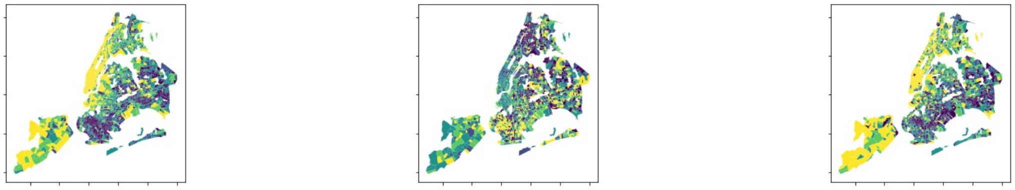


Figure 9: *Non-Twitter Features Correlation Heatmap*

3.2.3 Twitter Features

Structured

Given that the Twitter-using population is disproportionately whiter, more educated, and wealthier than the population as a whole (Wojcik and Hughes, 2019), differences that were particularly stark in the early 2010s, tweet growth in a particular census tract may indicate more visits from this demographic group, possibly heralding future or current moves that will spark gentrification. Furthermore, an increase in English-language tweets may also reflect gentrification, as non-English speakers, who tend to be less educated and have lower incomes than those whose main language is



(a) Avg Tweet Count (b) Percent English Trend (c) Tweet Count Trend

Figure 10: Structured Twitter Features by Census Tract

English (Cheeseman Day, J., & Shin, H.B., 2005), are crowded out by the latter group. From the Twitter dataset described in Section 2.2, the following structured features were constructed at census tract level

- Growth in the share of total tweets that are in English.
- Growth in the total count of tweets.
- Total count of tweets in the first time period in which someone tagged a location in that census tract in the data (usually but not always January of 2011).
- Total count of tweets in the last time period in which someone tagged a location in that census tract in the data (usually but not always December of 2013)
- Average number of monthly tweets between 2011-2013

The growth rates were calculated using the same line of best fit technique as the non-Twitter structured features. Any month-year-tract sets that did not contain geotagged tweets were included in the dataset as counts of 0 for that month before any structured features were calculated. Figure 10 demonstrates some of the extracted Twitter features, divided into quantiles. As before, lighter colours indicate higher values. Manhattan, inner Brooklyn, and Staten Island saw the highest number of average tweet counts; while southern Brooklyn and eastern Queens saw the fewest. Growth in tweets reflected average tweet count closely, while growth in the share of total tweets in English seemed to occur in areas with fewer average tweets, particularly southern Brooklyn.

Unstructured

I constructed unstructured data features from the text of the English-language tweets using the Natural Language Processing (NLP) technique of topic modelling, which attempts to extract a given number of topics from a set of documents – in this case, the tweets fed to the topic modeller. Topic modelling techniques rely on the simplifying assumption that the author of a text writes their text by choosing words from a series of baskets of words, or topics. Thus, any text should be able to be separated into the baskets, or topics, from which the words came (Graham et al., 2012). Each topic learned by the model can be represented by a set of keywords and a probability distribution indicating the likelihood that a given keyword appears in that topic. I used a Latent Dirichlet allocation (LDA) model, an unsupervised topic extraction model implemented by the gensim library in Python (Rehurek & Sojka, 2011) that groups words into a set of topics based on how often the words co-occur in a given document. LDA first determines a number of words associated with each topic, then outputs an array of scores corresponding to the probability that a given document can be represented by each topic.

Most topic modellers are used on much longer documents with more coherent topics. Tweets pose a unique challenge for topic modelling, given their extremely short (until 2017, tweets could not be more than 120 characters) and often disjointed text. I thus used the Mallet implementation (also in gensim) of LDA, which is considered particularly strong for modelling the topics of shorter texts. It uses Gibbs sampling, a Markov chain Monte Carlo algorithm, to estimate the topic probability distribution for each document in a corpus.

I also preprocessed the tweet data using gensim's simple preprocessing function, eliminating punctuation, emojis, capitalization, and accents. I removed all URLs, tags of other users (those words beginning with '@'), and the word '[pic]', which represents a picture posted by the user. New line characters ('') and stop words from nltk's English language stop word list were removed, and the data was lemmatized (converted to its root word, such as 'waking' or 'woke' to 'wake') using spaCy's (Honnibal & Montani, 2017) English language lemmatization model. Finally, each document was converted to a 'bag of words', which consists of a set of tuples representing each unique word in the corpus and how many times that word appears in a given document.

Coherence Scores

Potential topic models were evaluated using $u\!-\!mass$ coherence scores, which assess the quality of topics learned by the model by evaluating how often each of the topic keywords in a model co-occur. Specifically, the u-mass coherence is denoted by the following equation:

$$Coherence = \sum_{i<j} score_{u-mass}(w_i, w_j) \tag{2}$$

where

$$Score_{u-mass} = log\frac{D(w_i, w_j) + 1}{D(w_i)} \tag{3}$$

where $D(w_i, w_j)$ equals the number of documents in which words w_i and w_j appear together, while $D(w_i)$ represents the number of documents in which only word w_i appears, for all keywords w_i and w_j in each topic learned by the model. The coherence score for the entire model is represented by the mean of each topic's individual coherence score. Thus, higher coherence scores generally indicate a better model.

Collating and Filtering Tweet Documents

Because tweets are so short and have an asymmetric word frequency distribution (with the vast majority of words appearing only once and a few words appearing many times, even with stop words removed), I evaluated the topic modeller on several different versions of a randomly selected sample, *Individual Subsample*, which consisted of approximately 16% (or 707,346) of the tweets in the corpus of approximately 4.4 million English-language tweets, or the *Full Sample*.

For each corpus, I created a histogram of word frequency and evaluated the topic coherence scores for different numbers of topics. Ideally, coherence scores should first increase as the number of topics increases (as most corpuses should have more than one topic), then decrease once the number has reached its optimum point.

I first ran the topic modeller on *Individual Subsample's* corpus, whose documents consisted of the 707,346 individual tweets and 3.03 million words, 127,682 of which were unique. However, the brevity of each individual document made it difficult for the topic modeller to successfully model topics. Figure 11 demonstrates that coherence scores decreased as topic numbers increased.

The difficulties caused by the extreme conciseness of the documents in this corpus inspired the construction of a corpus with collated tweets, the *Collated Subsample.* To construct each document in this corpus, I combined the tweets for a given month-year-census tract (e.g., March 2011 in census tract 36085000900) from *Individual Subsample*, resulting in a maximum possible corpus length of 75,564 (12 months per year * 3 years * 2099 census tracts under consideration) documents. The finalized corpus consisted of 45,277 documents; while every census tract considered had at least one geotagged

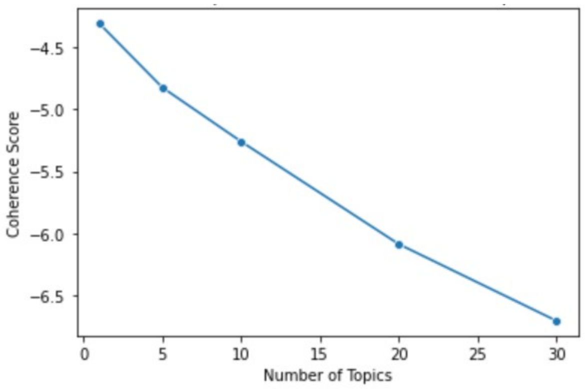


Figure 11: *Coherence Score vs. Number of Topics for Individual Tweets*

tweet between 2011-2013 in the full 4.4 million-tweet sample, the census tracts in the smaller *Collated Subsample* often did not contain geotagged tweets for every month.

Figure 12 demonstrates the coherence scores for a given number of topics between 0 and 200 in *Collated Subsample*; similarly to the *Individual Subsample*, coherence scores decrease as the number of topics increase.

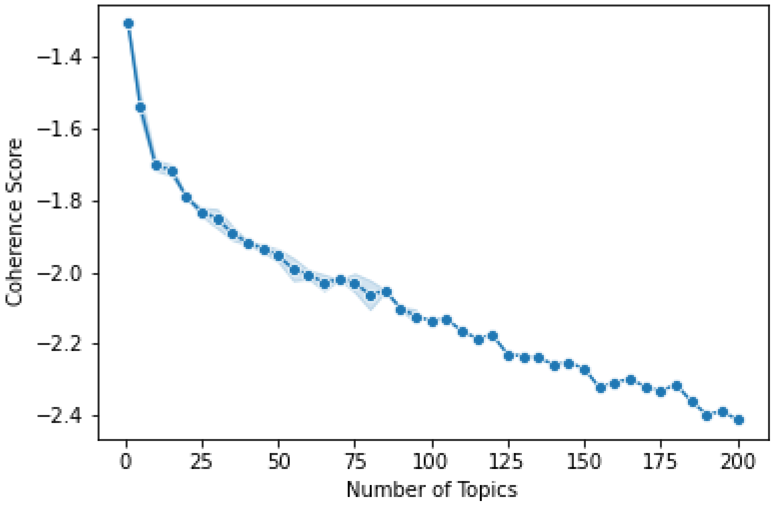


Figure 12: *Coherence Score vs. Number of Topics for Collated Tweets*

Figure 13 demonstrates the document frequency distribution; the y-axis corresponds to the number of words appearing in a given number of documents as represented by the x-axis. In the non-filtered *Collated Subsample*, a large number of words appear in only one document, and a small number of words appear in a very large number of documents. An examination of the frequency distribution reveals that 60% of all words appear in 1 document, 70% in 2 or fewer, and 80% in 3 or fewer.

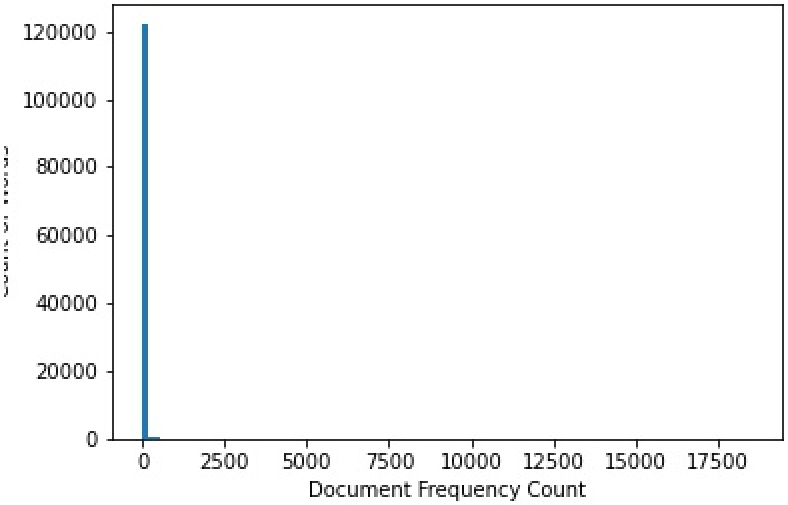
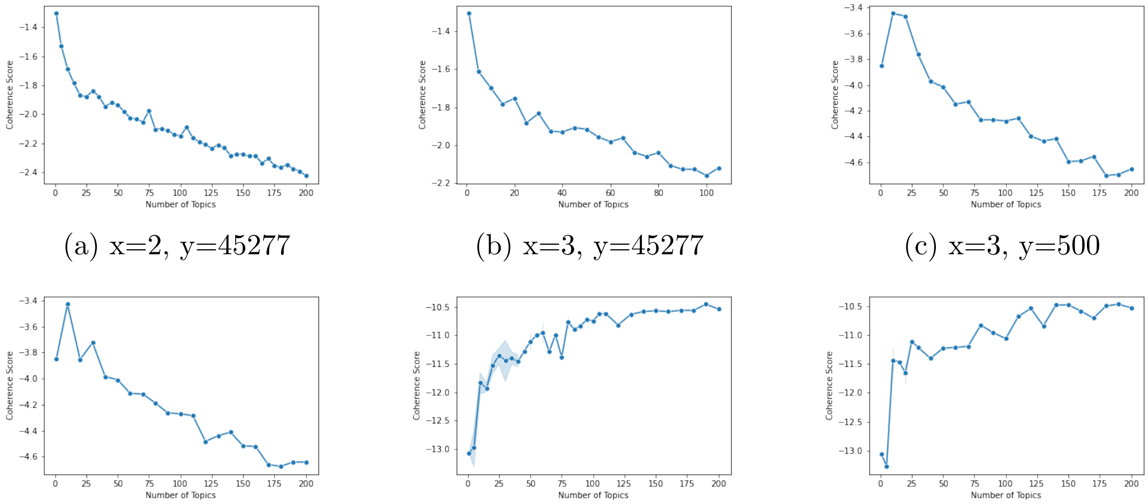


Figure 13: *Document Frequency for the Non-Filtered Collated Sample*

To improve the topic modeler, I thus filtered the corpus to create a less skewed document frequency distribution, dropping any words that appeared in fewer than x, where x = 0, 1, 2, 3, 4, or 5, documents, or more than y, where y = 100, 200, 500, or 45277 (the total number in the sample) documents. To test the quality of each filtered corpus, I created a histogram showcasing the document frequency distribution, as in Figure 13, and plotted the coherence scores for k topics, where k ranges from 0 to 200 in 5-unit increments. Figur 14 shows a selection of these plots, which follow an irregular monotonic pattern until words that appear in fewer than three documents (x=3) and more than 500 documents (y=500) are eliminated from the corpus, at which point coherence increases at first, then decreases as the number of topics grows.

Moreover, as demonstrated by Figure 15, a disproportionately large share of words (80%) appear in only 1, 2, or 3 documents in the *Collated Subsample*, while a very small number of words appear in over 500 documents.

Given the topic distribution patterns in the above figures, I filtered the corpus of the *Full Sample* by eliminating words that appear in fewer than four documents or more than 500. Figure 16 shows the document frequencies for the *Full Sample*, post-



(a) x=2, y=45277

(b) x=3, y=45277

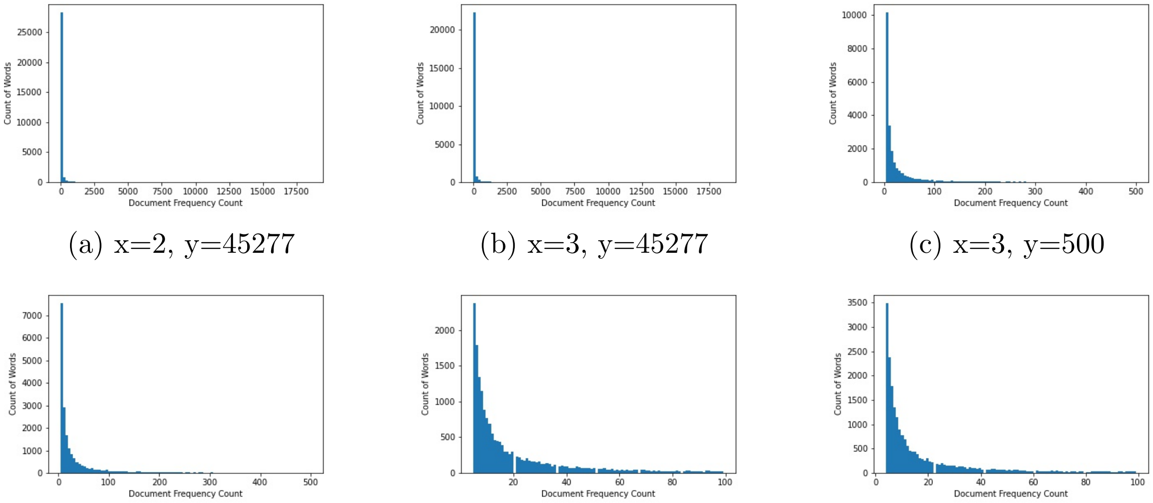
(c) x=3, y=500

(d) x=4, y=500

(e) x=3, y=100

(f) x=4, y=100

Figure 14: Filtered Corpus Coherence Scores on Collated Subsample



(a) x=2, y=45277

(b) x=3, y=45277

(c) x=3, y=500

(d) x=4, y=500

(e) x=4, y=100

(f) x=3, y=100

Figure 15: Filtered Corpus Document Frequency on Collated Subsample

filtering. The document frequency here broadly reflects that of the smaller samples seen in the previous Figures. The *Full Sample* originally consisted of 73,354 documents, with 438,895 unique words. The filtered sample still contains 73,354 documents, but only 72,179 unique words.

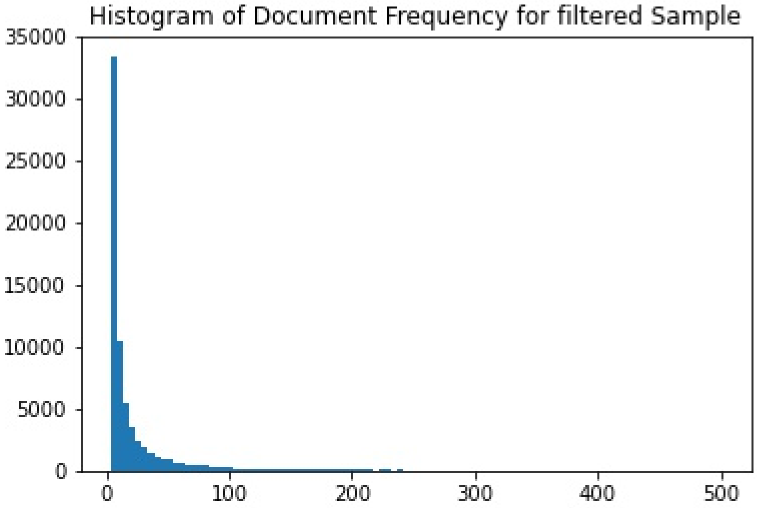


Figure 16: *Coherence Score vs. Number of Topics for Filtered and Collated Tweets*

To find the ideal number of topics for the topic modeller, I plotted the coherence scores for a given number of topics, ranging from 5 to 55 for the filtered sample, as seen in Figure 17, which revealed that the ideal number of topics is 45. Table 1 reveals some of the topics, described by ten keywords as well as the likelihood that each of those keywords appears in the given topic (following in parentheses), learned by the LDA Mallet topic model. Note that some words appear to be compilations of other words – those are hashtags (with the hashtag removed), while others appear to be shortened forms of a word, the result of lemmatization.

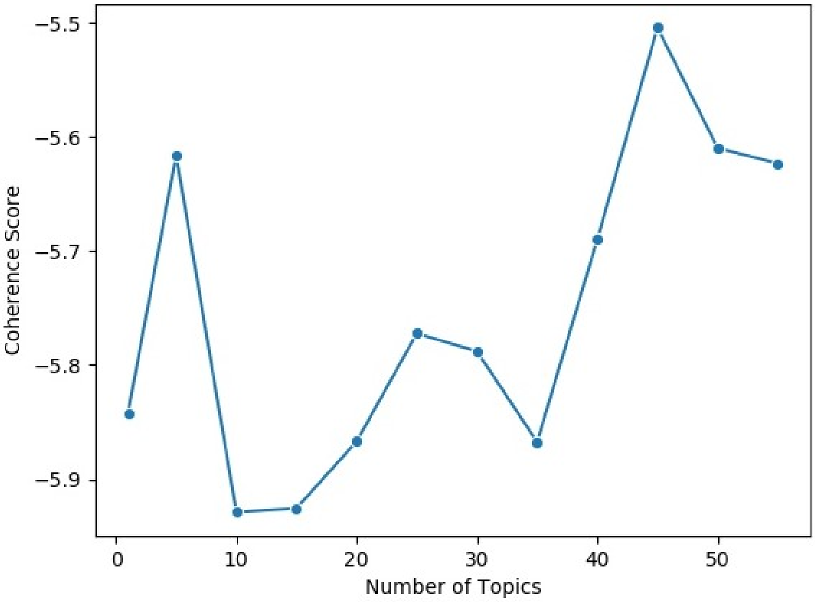


Figure 17: *Coherence Score vs. Number of Topics for Filtered and Collated Tweets*

Table 1: Select Topic Keywords and Probability

Topic Number	Keywords 1-10 (Probability)			
5	Jobcircle(0.109)	Analyst (0.048)	Trndnl (0.043)	Java (0.028)
	Sunnyside (0.019)	Misi (0.016)	Cybercoder (0.016)	Architect (0.016)
	Guidance (0.014)	Technical (0.013)		
22	Seaport (0.009)	Knot (0.006)	Beekman (0.005)	Obamacare (0.005)
	Relation (0.005)	Economic (0.005)	Journalism (0.004)	Voter (0.004)
	Provider (0.004)	Conservative (0.004)		
42	Assault (0.007)	Standtoendrape (0.007)	Gownaus (0.005)	Sexually (0.005)
	Roundup (0.005)	Thief (0.005)	NYPD (0.004)	Jury (0.004)
	Santo (0.004)	Firefighter (0.004)		

Choice of Best Topic Modelling Measure Once the ideal number of topics is chosen and the LDA Mallet model trained on the entire (filtered) sample of 73,354 documents, the model then returns a 45-element array for each document, with each element of the array representing the probability the document corresponds to one of the topics learned by the model. This results in a 73,354-by-45 element matrix; from this matrix, I extracted topic modelling features in three ways:

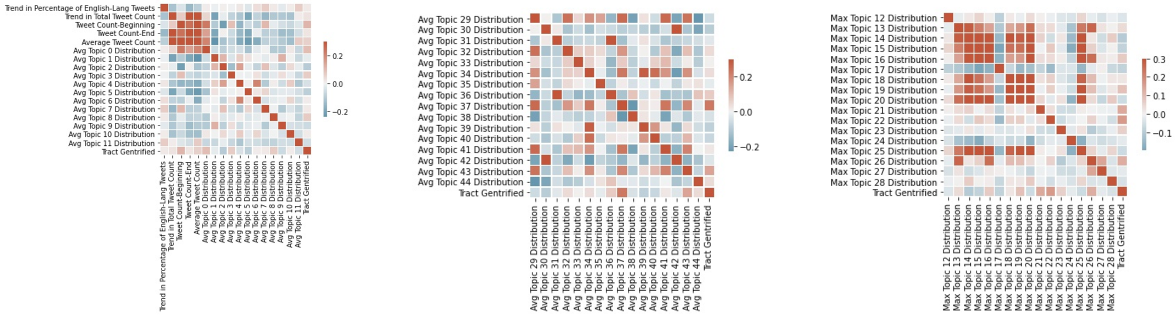
- The *mean* topic probability distribution for each topic and census tract, resulting in a 45-element vector for each tract. This represents the general state of (tweet) discussion in each census tract over the entire time period.

- The *maximum* topic probability distribution by topic, which is also a 45-element vector for each census tract. This represents the peak of discussion about any given topic, which could also prove effective in predicting gentrification from topics that are more irregularly discussed. For example, perhaps one-off events like food festivals are more likely to attract new residents that are young and well-educated; such events would likely be reflected by irregular discussion on Twitter better captured by maximum than mean topic probability distributions.

- The *temporal* topic probability distribution, which is represented by the topic probability vector for each month in the data set (resulting in a 45*36 = 1620 element vector for each tract). This allows the change in tweet topics over time to be represented in the model, which could prove more predictive than the other two sets of features. As one potential example, the topics in tweets from neighbourhoods in the process of gentrifying could become more similar to wealthier neighbourhoods over time, which may be reflected in the topic temporal distribution but not the mean or maximum distribution.

Figure 18 shows the correlation between some of these Twitter features (not all are shown due to space constraints; see the Appendix for others) and gentrification in eligible census tracts. Total tweet count at the start of the gentrification period (near January 2011) and many average or maximum topic distributions are correlated with gentrification.

To evaluate the relative quality of these measures, I used a cross-validation technique. The data was first randomly divided into an 85/15 train/test split, then normalized according to the mean and standard error of the entire set. The models were



(a) Mean Topic Corr Part 1 (b) Mean Topic Corr Part 3 (c) Max Topic Corr Part 2

Figure 18: Selected Topic Correlation Heatmaps

evaluated using the area under the receiver operating characteristic (ROC) curve (AUROC score) from 10-fold cross validation; they were trained on 90% of the training set, then evaluated on the other 10%. The process was repeated 10 times, so that every data point (apart from those in the test set) appeared in both the train and validation set.

I chose to use AUROC as my evaluation measure rather than a pure accuracy score to avoid the pitfalls that arise from the accuracy paradox. Because my classes are highly imbalanced, accuracy could be maximised by a trivial algorithm that simply predicts no gentrification; however, this would provide policy makers with no useful information whatsoever. The AUROC score calculates the area under the ROC curve, which plots the true-positive rate against the false-positive rate predicted by a model at differing thresholds; a perfect model would have an AUROC of 1, while a random classifier would have an AUROC of 0.5.

The various measures were tested against each other in the process of selecting the hyperparameters for the classification models (see Section 3.3 for more detail on the results), but the temporal topic probability distribution proved the best overall, with the highest AUROC scores under most conditions. Most, but not all census tracts contained geotagged tweets for every month in the gentrification period; I thus replaced any NA scores in the temporal tweet features with 0, which accurately reflects the zero probability that a particular topic could be linked to tweets in that month/census year combination (since they do not exist!).

For the full list of topic features, see Section 9.1 in the Appendix, which shows the features used in each of the models: baseline; baseline and non-Twitter features; baseline, non-Twitter behavioural, and Twitter features. The baseline model contains 6

Table 2: Select Summary Statistics

	Test	Train	All
Median Rent	950.703	924.711	928.754
Percent Black	0.397	0.429	0.424
Percent White	0.202	0.206	0.206
Percent Bachelors	0.153	0.163	0.161
Percent Unemployed	0.127	0.132	0.132
Percent of HHs Renting	0.867	0.845	0.848
Trend in Commercial Addresses	-0.007	0.026	0.021
Trend in Residential Addresses	0.349	0.493	0.47
Trend in ST Vacant Residential Addresses	-0.07	-0.028	-0.035
Trend in ST Vacant Commercial Addresses	0.032	0.013	0.016
Trend in Vacant LT Residential Addresses	-0.13	-0.113	-0.115
Trend in Vacant LT Commercial Addresses	0.027	0.004	0.008
Crime PC 06-10	0.007	0.008	0.008
Crime PC 11-13	0.007	0.007	0.007
Felonies PC 06-10	0.002	0.002	0.002
Felonies PC 11-13	0.002	0.002	0.002
Trend in Percentage of English-Lang Tweets	-0.297	-0.231	-0.242
Trend in Total Tweet Count	2.551	2.537	2.539
Tweet Count-Beginning	12.921	21.45	20.123
Tweet Count-End	100.159	119.178	116.22
Average Tweet Count	61.508	66.407	65.645

features, the baseline + non-Twitter model comprises 23 features, the full model consists of 1648 features, and the Twitter only model contains 1625 features. All socioeconomic features are taken from the 2006-2010 data (preceding onset of gentrification). Table 2 shows summary statistics for some of the features for the train set, the test set, and the full (combined) set.

3.3 Formulation of Classification Problem

Gentrification prediction is a binary classification problem; only eligible census tracts are considered, which then either gentrify or not over the time period considered. The features extracted the 2010 ACS data, non-Twitter data sources, and Twitter, are utilized to predict gentrification. Both logistic regression and random forest models, implemented with scikit-learn (sklearn; Pedgerosa et al.), were used to predict gentrification.

3.3.1 Logistic Regression (LR)

LR is a linear supervised machine learning technique that allows for binary prediction. Given feature vector $\mathbf{x}$, LR predicts class $y \in$ (-1,1), using the following equation, where Ber denotes the Bernoulli distribution:

$$p(y|\mathbf{x}, \mathbf{w}) = Ber(y|\frac{exp(\mathbf{w}^T\mathbf{x})}{exp(\mathbf{w}^T\mathbf{x} + 1})) \tag{4}$$

Where $\mathbf{w}$ denotes the weights given to vector $\mathbf{x}$, sometimes referred to as the Beta values (or β). To avoid overfitting, the model was regularised using ℓ_2 regularisation, also known as ridge regression, which adds a cost penalty to the optimisation of the equation to force the shrinkage of coefficients that are less important for the final classification. The following equation is minimised over $\mathbf{w}$:

$$f(\mathbf{w}) = (\sum_{i=1}^{N} log(1 + exp(-y_i w_i x_i)) + \lambda \mathbf{w}^T \mathbf{w} \tag{5}$$

The imbalance between positive and negative labels in the dataset (only about 24% of eligible census tracts gentrify) may pose challenges in classification. To overcome this challenge, in addition to the basic logistic regression model, I also evaluated several other versions, using the full model features:

Weighted model - In logistic regression, weights can be added to classes so that the effect of each weighted point is equivalent to x-times each unweighted point. I tested two version of a weighted model. The first used the weights that can be computed by sklearn's logistic regression model formula. This is approximately equivalent to the inverse class ratio; for example, if there are approximately 3 non-gentrified tracts (label = 0) for every gentrified tract, then the weights will be 1:3. In this case, the weights were computed as 0.6577:2.0853, which is equivalent to approximately 1:3.17. Using grid search, I also tested weights for the positive gentrification label ranging between 2.5 and 3.5 in 0.01-unit increments. To tune the weights, I used 10-fold cross-validation to evaluate which weight ratio returned the best AUROC score; that weight ratio was then chosen as the tuned weight.

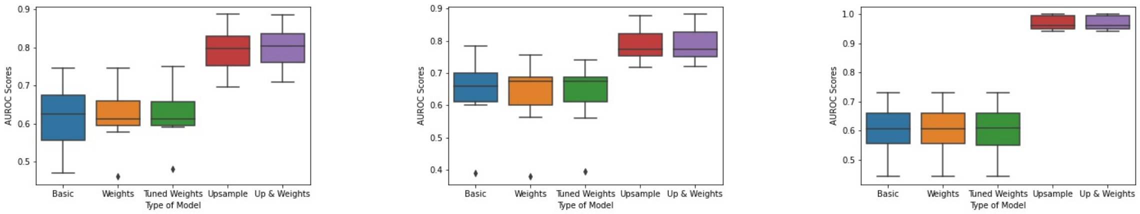
Upsampled model - One method to improve prediction models with an imbalanced dataset is to train the model on a balanced dataset created by bootstrapping (random resampling with replacement) the features of the class that occurs less frequently (in this case, the gentrified census tracts) – known as upsampling. This often improves prediction accuracy and the AUROC curve. I upsampled with a fixed random seed to ensure the reproducibility of my results. My original training set contained 260 non-gentrified and 82 gentrified census tracts; after upsampling, the new training set consisted of 260 gentrified as well as the original 260 non-gentrified census tracts.

Weighted and upsampled model- This method combines the previous two; I used cross-validation to find the best weights for the upsampled model, then used these weights with the upsampled dataset.

This resulted in five potential models: the basic model, the computed weights model, the tuned weights model, the upsampled model, and the upsampled and weighted model. To choose which Twitter model to evaluate against the baseline and non-Twitter models, I assessed the AUROC performance of these five models, testing each while using one of three topic distribution variables – maximum, mean, and temporal – as discussed in Section 3.2.3, which left me with 15 models in total.

To evaluate which of these 15 model setups performs the best, I divided my data into an 85/15 percent train/test split; each contained approximately the same number of positive (gentrified) labels, with 260 negative (non-gentrified) and 82 positive (23.9%) labels in the former and 48 negative and 15 positive (23.8%) in the latter. I then normalized the features in both the train and test sets using the overall set mean and standard error. I evaluated the AUROC performance using 10-fold cross-validation – training the data on 90% of the train set and calculating the AUROC score on the other 10% of the training set, then repeating this process until the entire set had been in the validation set. The mean of the 10 AUROC scores returned is used to evaluate the performance of the model.

Table 3 and Figure 19 compare the performance of all 15 models; clearly, upsampled models greatly outperform their non-upsampled counterparts. The logistic regression model with the best performance is the upsampled temporal topic distribution; I thus chose this model to evaluate against the baseline and non-Twitter models.



(a) Mean Topic Dist Models (b) Max Topic Dist Models (c) Temp Topic Dist Models

Figure 19: Comparing LR Topic Distribution Models

Table 3: Comparing AUROC Scores of LR Models

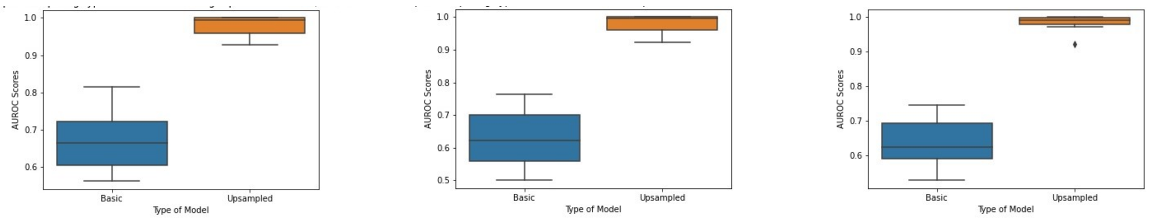
	Mean AUROC Score (SE)		
Type of LR Model	Mean Distribution	Max Distribution	Temporal Distribution
Basic	0.6171 (0.029)	0.6419 (0.033)	0.6011 (0.029)
Computed Weighted	0.6232 (0.025)	0.6356 (0.033)	0.6011 (0.029)
Tuned Weighted	0.6264 (0.024)	0.6358 (0.031)	0.6025 (0.03)
Upsampled	0.7919 (0.02)	0.7861 (0.016)	0.9691 (0.008)
Upsampled & Weighted	0.7979 (0.018)	0.7886 (0.016)	0.9692 (0.008)

3.3.2 Random Forest (RF)

RF is a non-linear supervised ML technique frequently used for classification problems. RF models create many decision trees that take the features at input and output a classification prediction; at every ‘node’ (branch in the tree), the model selects the feature that best allows it to split the data into two smaller datasets, ideally as dissimilar as possible. The dissimilarity of the resulting datasets is known as that node’s ‘purity’ – if all the examples in each smaller dataset belong to a separate class, the node is 100% pure; if it is split 50/50, then the node is 0% pure (or 100% impure).

Each of these datasets is a ‘leaf’. This procedure continues until the tree reaches a maximum depth (number of splits) or the leaf can no longer be split. To minimise overfitting, RF models train a multitude of decision trees on a random subset of bootstrapped (random sampling with replacement) observations from the training set, then ‘vote’ on the final classification of each observation by averaging the classifications that each model returns (Reades et al., 2018).

To return the model with the best performance, I first tuned the hyperparameter of number of estimators, or the number of trees in the forest, which can significantly change the performance of a RF model. I used the same 85/15 train/test split as when evaluating the Logistic Regression model. To tune this hyperparameter, I used ran-



(a) Mean Topic Dist Models (b) Max Topic Dist Models (c) Temp Topic Dist Models

Figure 20: Comparing RF Topic Distribution Models

Table 4: Comparing AUROC Scores of RF Models

	Mean AUROC Score (SE)		
Type of RF Model	Mean Distribution	Max Distribution	Temporal Distribution
Basic	0.6699 (0.027)	0.6275 (0.028)	0.6345 (0.022)
Upsampled	0.9791 (0.008)	0.9778 (0.009)	0.9834 (0.007)

domized search cross-validation, considering a random selection of estimators between 10 and 2000 and returning the estimator with the highest AUROC score for each of six potential models, described below.

Once the hyperparameter was tuned, I then evaluated six models with the same tenfold cross validation method used for logistic regression: a basic random forest model versus an upsampled model for each of the three topic distribution measures. Figure 20 and Table 4 show the results of this evaluation; like the LR models, the RF upsampled datasets greatly outperformed their imbalanced counterparts, while the features extracted from the temporal topic distribution slightly outperforms the mean and maximum topic distribution features. Thus, the upsampled temporal topic distribution model was chosen to be evaluated against the baseline and non-Twitter models.

3.4 Feature Importance

Finally, I sought to understand the importance of different features to the model. Understanding which features the models relied most heavily on can shed light on the process and causes of gentrification. To investigate feature importance, I used two measures:

Logistic Regression Weights: The weights (or Beta values) from the trained model indicate which features are the most important; the higher the absolute value of the feature, the more impact it has on the final classification.

Random Forest Gini Importance: The Gini importance measures the mean decrease in 'impurity' each feature provides. As a reminder, a node in a decision tree is impure if it fails to split a dataset into distinct sets whose classifications do not overlap. More useful features should be better able to split a mixed-classification dataset into pure single class nodes. For instance, if all census tracts with a median household income above $40,000 gentrified, while all those below this threshold did not, this feature could be easily used to split and accurately predict a dataset with both types of census tracts. More specifically, a feature's Gini importance is defined as the total decrease in node impurity weighed by the proportion of all samples reaching that node and divided by the total number of trees in the model.

4 Results

4.1 Model Performance

I compared the performance of the three models – the baseline, baseline and non-Twitter, and the full model (including baseline, non-Twitter, and Twitter features). Since training the model on an upsampled train set always improved the performance on the test set, I present only the results of the models trained on the upsampled dataset; the same dataset is used for all three models. The Appendix includes the results of the baseline and Twitter models trained on the non-upsampled dataset, as well as models containing only Twitter features; the results broadly reflect those in this section. To evaluate model performance, I calculated the AUROC score for each of the trained models on the held-out test set, which was normalized with the mean and standard deviation of the training set to avoid any leakage from the test set to the train set. Importantly, although the models are trained on an upsampled dataset, the test set involves only the original, non-resampled dataset. The test set, which contains 63 observations in total, includes 48 negative (non-gentrified) labels and 15 positive (gentrified) labels (or approximately 23.8% of the total). Table 5 compares the AUROC score performance of the three models on the held-out test set. Overall, the full Twitter model performs the best, followed by the non-Twitter model and then the baseline model, for both the Logistic Regression (LR) and Random Forest (RF) models. A perfectly random classifier would return an AUROC score of 0.5; thus, all the models except for the RF baseline perform better than random. Overall, the full Twitter LR model performs the best, with an AUROC score of 0.68, much higher than its baseline and non-Twitter counterparts (0.53 and 0.61 respectively).

Table 5: Comparing Test Set AUROC Performance of Upsampled Models

	Baseline Model	Non-Twitter Model	Full Twitter Model
Logistic Regression	0.5306	0.6069	0.6819
Random Forest	0.4785	0.6410	0.6597

Baseline Upsampled Model Confusion Matrix

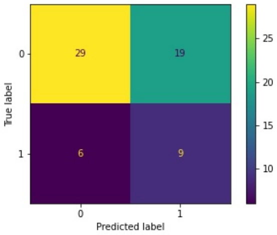


Figure 21: *Logistic Regression Baseline*

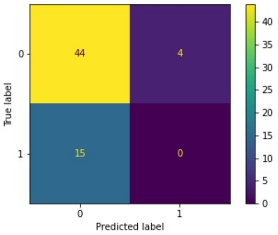


Figure 22: *Random Forest Baseline*

Non-Twitter Upsampled Model Confusion Matrix

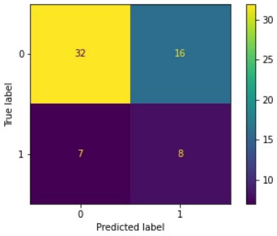


Figure 23: *Logistic Regression Non-Twitter*

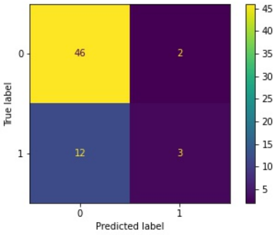


Figure 24: *Random Forest Non-Twitter*

Full Twitter Upsampled Model Confusion Matrix

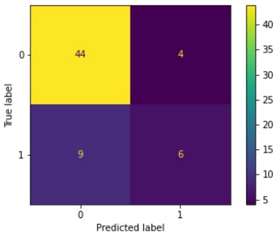


Figure 25: *Logistic Regression Twiiter*

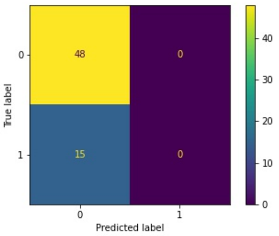


Figure 26: *Random Forest Twitter*

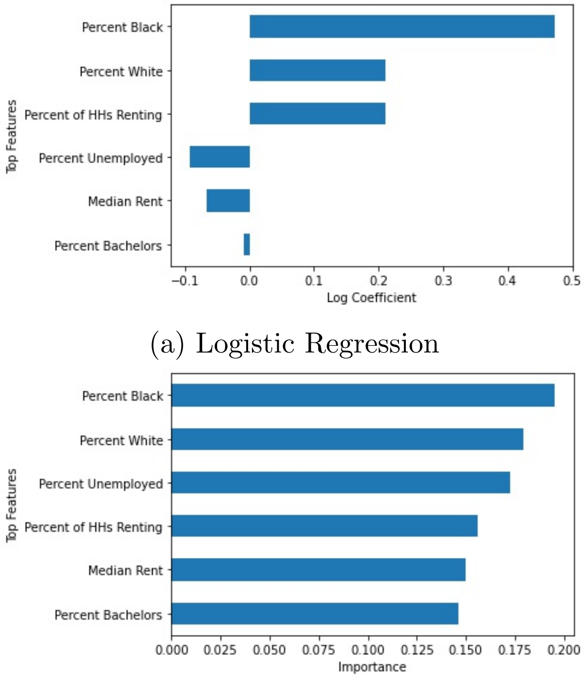
Figures 21-26 plot the confusion matrices of each model, showing the number of true and false positives as well as true and false negatives. The LR models outperform their Random Forest counterparts in both the baseline and full model; in both cases, the RF model fails to accurately predict any true positives, perhaps due to potential overfitting.

The baseline and non-Twitter LR models predict the most true positives, but they also predict many false positives, with a lower precision (32% and 33%, respectively) than the full model (60%). However, the baseline model's recall score (60%) is higher than either of the other two models (53% and 40% for the non-Twitter and full Twitter models, respectively).

In the RF models, precision is equally low for the baseline and full Twitter models (0% for both), and higher for the non-Twitter model (60%), while the non-Twitter model has the highest recall (20% versus 0% for both the other two).

4.2 Feature Importance

Figures 27, 28, and 29 reveal the most relevant features for each model, using the feature weights from the LR and Gini importance from the RF models as a measure of importance. While the feature importance is reported for all six features in the baseline model, only the top 15 features in the other two models are reported given the much greater number of features. For the exact values, see Tables 9 and 10 in the appendix. All Gini importance measures are positive, but many LR weights are negative, indicating that feature negatively impacted the probability of a positive label. However, both sets of features are ordered by their absolute value, so that highly negative values are counted along with highly positive ones. For example, crime per-capita between 2006-2010 is the second most important feature in the non-Twitter model, but it is negative, indicating that higher per-capita 2006-2010 crime rates are negatively associated with



(a) Logistic Regression

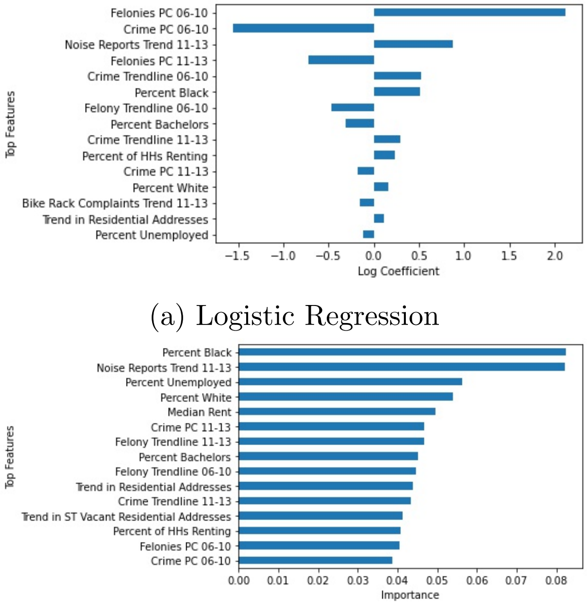
(b) Random Forest

Figure 27: Baseline Model Feature Importance

the later gentrification of a census tract.

The LR and RF models report similar feature importance rankings for the baseline and non-Twitter models; in the baseline model, race is the most important feature for both, while percent unemployed is the least. Although the order is not identical, the LR and RF non-Twitter models both share all but three of the same features in their top 15, including overall and crime per-capita trends, noise reports, percent university-educated, and percent of households renting.

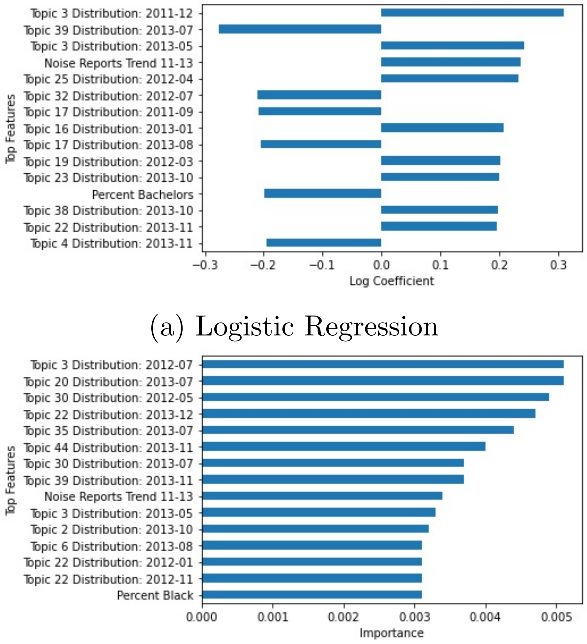
The figures indicate that the features extracted from the unstructured topic modelling are the most important; in the full model, nearly all the top features fall under this category. Tables 6 and 7 show the keywords associated with the top topics (the full list can be found in the Appendix). Some keywords are combinations of other words – those are hashtags, a common metadata tag used on Twitter to indicate the topic of the tweet, although the hashtag itself has been removed. Because the topics are temporal, a few – topics 3, 17, 30, and 22 – appear multiple times in the same model. Notably, all the topics with the greatest importance in the RF model, and nearly all in the LR model, are more recent (from 2012 and 2013 rather than 2011).



(a) Logistic Regression

(b) Random Forest

Figure 28: Non-Twitter Behavioural Model Feature Importance



(a) Logistic Regression

(b) Random Forest

Figure 29: Full Twitter Model Feature Importance

Table 6: LR Full Model Top Topic Keywords: In Order of Importance

Topic	Coeff Direction	Keywords
3	Positive	beacon, apr, loew, nostrand, pourhouse, facility, athletic, checkout, pickup, kingdom
39	Negative	arena, knickstape, kew, bally, pacer, isle, bruin, okc, islander, clipper
25	Positive	biergarten, meatpacke, gansevoort, meatpacking, snowpocalypse, hum, brass
25 (cont)	Positive	smorgasburg, vicious, cameo
32	Negative	timessquare, regal, imax, bubba, gump, tussaud, potter, avenger, foxwood, rpx
17	Negative	teamfollowback, followback, wht, billiard, cnt, nf, williamsbridge, pelham, bak, thnx
16	Positive	greenmarket, handmade, caroline, tkts, assembly, sandal, specifically, tapa
16 (cont)	Positive	unionsquare, triangle
19	Positive	parson, narrow, knit, mercer, riverdale, toll, ribbon, roeble, villagevine, spuyten
23	Positive	getalljobs, tinyurl, accounting, sporting, equity, ecommerce, nursing, analyst
23 (cont)	Positive	healthcare, employment
38	Positive	tide, concourse, rod, gowanus, redsox, oriole, dodger, loll, sanchez, nyy
22	Positive	seaport, knot, beekman, obamacare, relation, economic, journalism, voter,
22 (cont)	Positive	provider, conservative
4	Negative	blockhead, sheepshead, canal, balloon, coal, veteran, ritz, cafeteria, lombardi, thrift

Table 7: RF Full Model Top Topic Keywords: In Order of Importance

Topic #	Keywords
3	beacon, apr, loew, nostrand, pourhouse, facility, athletic, checkout, pickup, kingdom
20	flatiron, kindle, ebook, tube, dekalb, barbecue, auditorium, nightclub, ctr, carnegie
30	mph, humidity, temperature, meadow, expwy, manor, vanderbilt, steady, springfield, utopia
22	seaport, knot, beekman, obamacare, relation, economic, journalism, voter,
22 (cont)	provider, conservative
35	northbnd, bnd, rv, nycha, armory, bypass, fabipop, satisfied, canal, whitehall
44	junction, thot, trill, foh, kendrick, disrespectful, bae, idgaf, nvm, matchless
39	arena, knickstape, kew, bally, pacer, isle, bruin, okc, islander, clipper
2	woodside, ditmar, pisce, briarwood, youuu, montauk, soundtracke, youuuu, ik, regent
6	essex, roaster, ludlow, eventsinnewyork, dumple, splash, tropical, rockwood,
6 (cont)	stumptown, baking

Many of the keywords are associated with places or neighbourhoods, including hotels (Loew, Ganesvoort, Ritz), affluent neighbourhoods (Flatiron, Montauk, Meatpacking [District], Times Square, Union Square, Riverdale, Williamsbridge, Spuyten, Seaport, Sheepshead [Bay]), gentrifying neighbourhoods (Gownaus), towns (Pelham), casinos (Foxwood), parks (Pelham), streets and/or subway stops (Nostrand), and universities (Parson for Parsons School of Design). Many others can be linked to sports, such as Islander, Bruin, Pacer, Clipper, Red Sox, Oriole, Dodger, NYY (New York Yankees). Others are political (assembly, Obamacare, journalism, voter, conservative) or economic (getalljobs, accounting, sporting, equity, nursing, ecommerce, employment). Some relate to restaurants – bubba, gump – or other commercial entreprises: Regal (a chain of cinemas), IMAX, Tussad, Smorgasburg (an open-air food market in Williamsburg), and Biergaten (the German word for beer garden, and part of the name of several NYC beer gardens). Some are slang words: loll, ik (short for ‘I know’), wht (short for ‘what’), foh (f*** outta here), bae (before all else, a slang word for a romantic partner), idgaf (I don’t give a f***), nvm (never mind), and thot (a term for a promiscuous woman).

Overall, keywords from the LR model topics with a positive coefficient are associated with baseball, economics and employment, politics, and neighbourhoods. Those with a negative coefficient are associated with basketball and hockey, certain types of commercial enterprises (the cinema, Madame Tussad's), or slang/shortened forms of words such as 'what' and 'thanks.' Because RF feature importance is only positive, only the relative magnitude of topic importance is relevant; the most important topics have keywords associated with politics, slang or vulgar words, sports and affluent neighbourhoods outside of Manhattan.

5 Discussion

5.1 Insights from Non-Twitter Features

In the baseline model, race and percentage of households renting are positively associated with gentrification, while percent unemployed and median rent is negatively associated with a positive label, implying that neighbourhoods with more renters, more white and black residents, fewer unemployed residents, and lower rent are more likely to gentrify. The association between share of households renting, unemployment, and gentrification seem consistent with theory and past work on the subject. One theory of gentrification, the 'rent gap', holds that gentrification occurs when the disparity between the current rental income of a property and the achievable income grows large (Clark and Gullberg, 1997; Lees et al., 2008; Smith, 1979; Smith & DeFillipis, 1999); a higher proportion of renting households could mean that the rent-gap incentive to raise the rent and attract highly educated, wealthier individuals is greater than in a neighbourhood with more owner-occupiers, thus leading to a greater probability of gentrification. The relationship between the share of residents that are black, starting rent, and gentrification seems inconsistent; this could be because once other variables are considered, more black residents is a stand-in for other variables that lead to gentrification, or it could be an indication of imperfection in the model.

In the non-Twitter models, crime, race, and noise complaints trends seem to be influential in the results of both models. The trend in noise complaints is also the only variable to show up in the top 15 most important features in both the LR and RF full models, which is consistent with previous work indicating that an increase in noise complaints may indicate a wave of new, young, educated residents moving into a neighbourhood and thus serve as a warning sign of gentrification. (Thuy Vo, 2018;

Legewie & Schaeffer, 2016).

5.2 Role of Twitter

The improvement in AUROC scores and the relatively high importance of Twitter features in the full model indicate that Twitter can be a useful tool for predicting as well as understanding gentrification. This is consistent with previous research and policy work on this subject; past studies have successfully used social media data to improve gentrification prediction and mobility, while policymakers' attempts to build gentrification prediction models from mainly socioeconomic variables have failed for the most part (Poorthuis et al., 2021; Jain et al., 2021; Cagney et al., 2020; Phillips et al., 2019; Glaeser et al., 2018; Chapple and Zuk, 2016; Preis et al., 2021).

The results from the full models may also help shed light on gentrification in New York City. For example, tweets about affluent neighbourhoods associated with gentrification may indicate increased mobility patterns between affluent neighbourhoods and poorer neighbourhoods about to gentrify; this could be the result of residents of affluent neighbourhoods traveling to poorer neighbourhoods, or vice versa, but without further information on the users, it is difficult to say. These results are consistent with previous work on gentrification (Poorthuis et al., 2021) as well as segregation (Phillips et al., 2019).

The positive association between politics (from words such as 'Obamacare', 'Assembly', 'voter', and 'conservative') and gentrification may reflect a change in neighbourhood composition – more educated individuals, whose movement into low-income neighbourhoods drives gentrification, are more likely to participate in politics (Persson 2015) – or an inherent link between politics and gentrification. Perhaps residents of a gentrifying tract are more driven to discuss politics due to their experiences. Further information about those tweeting about politics – where they live or information about their background, which could be gleaned from their Twitter bio, the text of their other tweets, and other geotagging, could help disentangle these possible effects. For instance, if they often geotag from the same location between 9-5 on weekdays, this is likely their workplace, while a frequent nightly and weekend location may indicate their home. In this way, a map of individuals' movements could be reconstructed, which could aid in answering this question. The results also indicated a link between discussion of employment (with words like 'accounting', 'nursing', 'ecommerce', and 'employment') or economics ('equity') and gentrification. This could indicate a rise in employment in census tracts that later gentrify or a change in the composition of the neighbourhood – perhaps those who tweet more about economics or employment are moving to formerly low-income neighbourhoods and triggering gentrification in the process.

The negative link between slang words and gentrification is also interesting. A further link between slang or vulgar words was observed in the RF model results, although the direction of the link is unknown. This result may reflect differences in word usage between residents of non-gentrifying and gentrifying neighbourhoods; given that more educated people, who may be less likely to use slang or vulgar words in tweets, often drive gentrification, this differential may be reflecting moves from such individuals into neighbourhoods that gentrify.

6 Conclusion

6.1 Summary

Gentrification poses a significant problem for policymakers today. Despite many efforts to create a gentrification forecaster, success in this arena remains elusive. Some recent work has successfully integrated social media to improve gentrification prediction (Glaeser et al., 2018; Jain et al., 2021; Poorthuis et al., 2021). This study aimed to build on previous work by using data from Twitter and other sources accessible to policymakers, including 311 complaints, crime rates, and address vacancies or changes, in addition to common socioeconomic features used in previous attempts, to predict which census tracts would gentrify from a set of 405 eligible tracts in NYC between 2010-2018. From Twitter, I extracted both structured and unstructured features. Using a Mallet LDA topic modeller, I estimated the probability that the tweets in a given month-year-census tract matched the topics learned by the model. I then built random forest and logistic regression models that were tuned using tenfold cross-validation to predict gentrification. Both the non-Twitter and full models (including Twitter features) improved upon the baseline model, returning AUROC scores of up to 0.68, much better than random chance. An investigation of the importance of the features in the model also revealed links between gentrification and certain topics, such as employment, politics, sports, or the use of slang words.

6.2 Contribution of Work

This model was intended as a 'proof-of-concept' for the use of Twitter in studying gentrification; it successfully integrated Twitter features to improve upon both the baseline and non-Twitter model. Moreover, given that Twitter is relatively more accessible than many other social media sites, policymakers are relatively more able to use it in their work. Thus, the integration of publicly available non-Twitter along with the Twitter data showed the possibilities for policymakers of using accessible data to aid in gentrification prediction. Along with the current contributions in the literature, this study has shown the potential power of social media data in predicting gentrification.

6.3 Limitations and Future Research

Like any study, this work faced several limitations. Firstly, the trained model was only tested on a single held-out test set. The scores might be a result of random chance rather than improved predictive power; future work should re-shuffle the train/test splits, then re-tune and re-train the models to see if the test scores differ significantly.

Furthermore, while NYC is often studied in gentrification studies, as the largest and densest city in America, it is unique in many ways that could impact the generalizability of the results to other cities. Moreover, little research on gentrification has been conducted outside North America and London; future research would do well to expand not only to other American cities, but also cities in other parts of the world that are less studied.

Tweets are a particularly difficult set of documents to perform topic modelling on, due to their often irregular spellings and extremely short text. Although I attempted to mitigate this by using a model specially designed for shorter text and extensively cleaning and lemmatizing the data, the cleaning was not perfect; for instance, one topic included the words 'youuu' and 'youu', which are both slang spellings of 'you.' Of course, the use of slang or shortened words (as separate from the full words) in and of themselves might be useful knowledge; for example, their use may indicate a certain level of education or occupation.

Moreover, the results could be sensitive to the exact definition of gentrification, especially given how few census tracts were considered 'gentrified' under the definition that I used. This imbalance can pose difficulties for the models. Most studies on gentrification rely on only one definition, but future research should test several different definitions to see if model results remain consistent.

Finally, though this work may shed light on the process of gentrification, rather than simple association, causality is nearly impossible to assert given the many biases present in the data, particularly selection bias. As more data becomes available, especially on individuals' movements from GPS phone tracking, research on individual users can help mitigate many of the questions raised. Moreover, much more research could be performed using the Twitter data. Beyond topic modelling, several other NLP techniques, such as doc2vec (a numerical representation of words) or sentiment analysis, are promising.

7 Acknowledgements

I am grateful to my family and friends, especially Angela Hosbein, Angelique Talmor, Patrick Savoie, Nils Bohr, and Robin Park, for being pillars of support during this incredibly difficult year. Thanks to Freddie Hoareau for cooking me many meals and letting me work on this thesis at his place for hours on end. I'm grateful to all the heroes on Stack Exchange who answered so many of my questions without me having to even ask them. Thanks to Drs. Philip Kalikman, Maxim Pinkvoskiy, Aaron Peterson, and Michelle Shumate, whose letters of recommendation or advice helped me attend Oxford. I'm grateful for my professors at the OII, who taught me so much despite the difficulties of this year. Finally, I would like to thank my advisors, Xiaowen Dong and Scott Hale, for the countless hours of hard work they put in to help me. They're two of the best advisors I've ever had, and I couldn't have done this without them.

8 References

311 service requests from 2010 to present: New York City. (2021). NYC OpenData. https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-\2010-to-Present/erm2-nwe9.

Academic Research Product Tract. (2021). Twitter. https://developer.twitter.com/en/products/twitter-api/academic-research.

Atkinson, R. (2000a). Measuring gentrification and displacement in greater London. *Urban Studies*, 37(1), 149–165. https://doi.org/10.1080/0042098002339.

Atkinson, R. (2000b). The hidden costs of gentrification: displacement in central London. *Journal of Housing and the Built Environment* 15 (4), 307-326.

Atkinson, R. (2003). Misunderstood saviour or vengeful wrecker? The many meanings and problems of gentrification. *Urban Studies*, 40(12), 2343–2350. https://doi.org/10.1080/0042098032000136093.

Boy, J. D., & Uitermark, J. (2016). How to study the city on instagram. *PLOS ONE*, 11(6). https://doi.org/10.1371/journal.pone.0158161.

Cagney, K. A., York Cornwell, E., Goldman, A. W., & Cai, L. (2020). Urban mobility and activity space. *Annual Review of Sociology*, 46(1), 623–648. https://doi.org/10.1146/annurev-soc-121919-054848.

Chapple, K. (2009.) Mapping susceptibility to gentrification: the early warning toolkit. Berkeley, CA: The Center for Community Innovation. http://communityinnovation.berkeley.edu/reports/Gentrification-Report.pdf.

Chapple, K., & Zuk, M. (2016). Forewarned: The use of neighborhood early warning systems for gentrification and displacement. *Cityscape: A Journal of Policy Development and Research*, 18(3), 109–130. https://www.jstor.org/stable/26328275.

Cheeseman Day, J., & Shin, H.B. (2005). *How does ability to speak English affect earnings?* [Paper presentation]. Annual Meetings of the Population Association of America: Philadelphia, PA. https://www.census.gov/content/dam/Census/library/working-papers/2005/demo/2005-Day-Shin.pdf.

Chong, E. (2017, Sept. 17). Examining the negative impacts of gentrification. *George-*

town Journal on Poverty Law & Policy. https://www.law.georgetown.edu/poverty-journal/blog/examining-the-negative-impacts-of-gentrification/.

Clark, E. (2005). The order and simplicity of gentrification: A political challenge. In: R. Atkinsonand G. Bridge (Eds.) *Gentrification in a global context: The new urban colonialism.* London: Routledge, pp. 261–269.

Clark, E. & Gullberg, A. (1997). Power struggles in the making and taking of rent gaps: The transformation of Stockholm city. In: O Kalltorp, I Elander, O Ericsson, & M Franzen (Eds,), *Cities in transformation – transformation in cities: Social and symbolic change of urban space.* Aldershot: Avebury, pp. 248–265.

CPI Inflation Calculator. *U.S. Bureau of Labor Statistics.* https://www.bls.gov/data/inflation_calculator.htm.

Cranshaw, J., Schwartz, R., Hong, J.I., & Sadeh, N. (2012). The livehoods project: utilizing social media to understand the dynamics of a city. *ICWSM.*

Cycling in New York City, 2007 to 2014. (2016). New York City Department of Health and Mental Hygiene. https://www1.nyc.gov/assets/doh/downloads/pdf/epi/databrief78.pdf.

Davidson, M. & Lees, L. (2005). New-build 'gentrification' and London's riverside renaissance. *Environment and Planning A: Economy and Space* 37(7): 1165–1190.

Ding, L., Hwang, J. & Divringi, E. (2016) Gentrification and residential mobility in Philadelphia. *Regional Science and Urban Economics* 61: 38–51.

Dragan, K., Ellen, I. G., & Glied, S. (2020). Does gentrification displace poor children and their families? New evidence from Medicaid data in New York City.*Regional Science and Urban Economics*, 83, 103481. https://doi.org/10.1016/j.regsciurbeco.2019.103481

Edlund, L., Machado, C., & Sviatschi, M. M. (2016). Bright minds, big rent: gentrification and the rising returns to skill. *SSRN Electronic Journal.* Published. https://doi.org/10.2139/ssrn.2871597

Florida, R. (2002, May). *The rise of the creative class.* Washington Monthly. https://washingtonmonthly.com/magazine/may-2002/the-rise-of-the-creative-class/

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

Gardiner, O., & Dong, X. (2021, Jan 5). Mobility networks for predicting gentrification. In: Benito R.M., Cherifi C., Cherifi H., Moro E., Rocha L.M., Sales-Pardo M. (Eds), Complex Networks & Their Applications IX. COMPLEX NETWORKS 2020 2020. *Studies in Computational Intelligence*, vol 944. Springer, Cham. https://doi.org/10.1007/978-3-030-65351-4_15.

Glaeser, E. L., Kim, H., & Luca, M. (2018). Nowcasting gentrification: using yelp data to quantify neighborhood change. *SSRN Electronic Journal.* Published. https://doi.org/10.2139/ssrn.3123733

Graham, S., Weingart, S., & Milligan, I. (2012). *Getting started with topic modeling and MALLET.* The Programming Historian. https://tinyurl.com/mxwasu5f.

Gould Ellen, I., Mertens Horn, K., & Reed, D. (2019). Has falling crime invited gentrification? *Journal of Housing Economics* 46 (101636). https://doi.org/10.1016/j.jhe.2019.101636/

Honnibal, M., Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

HUD 2010 income limits. Department of Housing and Urban Development Office of Policy Development and Research. https://www.huduser.gov/portal/datasets/il.html#2010.

Jain, S., Proserpio, D., Quattrone, G., & Quercia, D. (2021). *Nowcasting gentrification using Airbnb data* (Vol. 5, Issue CSCW1). Association for Computing Machinery (ACM). https://doi.org/10.1145/3449112.

Lees L., Slater, T. & Wyly, EK (2008) *Gentrification.* New York: Routledge.

Legewie, J., & Schaeffer, M. (2016). Contested boundaries: explaining where ethnoracial diversity provokes neighborhood conflict. *American Journal of Sociology* 122 (1): 125-161.

Lung-Aman, W. (2021, May 19). *Small businesses are victims of gentrification, too.* Bloomberg City Lab. https://www.bloomberg.com/news/articles/2021-05-19/small-businesses-are-victims-of-gentrification-too.

Manson, S., Schroeder, J., Van Riper, D., Kugler, T., & Ruggles, S. IPUMS National Historical Geographic Information System: Version 15.0 [dataset]. Minneapolis, MN: IPUMS. 2020. http://doi.org/10.18128/D050.V15.0.

NYC 311. NY State: Local and Regional Authorities. https://www.ny.gov/agencies/nyc-311.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *JMLR* 12: 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

Persson, M. Education and political participation. (2015) *British Journal of Political Science* 45 (3): 689–703. DOI:https://doi.org/10.1017/S0007123413000409.

Phillips, N. E., Levy, B. L., Sampson, R. J., Small, M. L., & Wang, R. Q. (2019). The social integration of American cities: Network measures of connectedness based on everyday mobility across neighborhoods. *Sociological Methods & Research*, 50(3), 1110–1149. https://doi.org/10.1177/0049124119852386.

Poorthuis, A., Shelton, T., & Zook, M. (2021). Changing neighborhoods, shifting connections: mapping relational geographies of gentrification using social media data. *Urban Geography*, 1–24. https://doi.org/10.1080/02723638.2021.1888016.

Porter, J. (2019, Jun 19). *Twitter removes support for precise geotagging because no one uses it.* The Verge. https://www.theverge.com/2019/6/19/18691174/twitter-location-tagging-geotagging-discontinued-removal

Preis, B., Janakiraman, A., Bob, A., & Steil, J. (2020). Mapping gentrification and displacement pressure: An exploration of four distinct methodologies. *Urban Studies*, 58(2), 405–424. https://doi.org/10.1177/0042098020903011.

Prince, S. 2014. *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. (2018). Understanding urban gentrification through machine learning. *Urban Studies*, 56(5), 922–942. https://doi.org/10.1177/0042098018789054.

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).

Rotondaro, V. (2019, Oct. 23). *Is it time for American cities to stop growing?* Vox. https://tinyurl.com/y6ckes8a.

Rubenstein, D., & Cheney, B. (2015, May 5). *Uber biggest in Manhattan and gentrified Brooklyn and Queens.* Politico NY. https://tinyurl.com/462bb58d.

Smith, N. (1979). Toward a theory of gentrification: A back to the city movement by capital, not people. *Journal of the American Planning Association* 45(4): 538–548.

Smith, N. & DeFilippis, J. (1999). The reassertion of economics: 1990s gentrification in the Lower East Side. *International Journal of Urban and Regional Research* 23(4): 638–653.

Thuy Vo, L. (2018, June 19). *They played dominos outside their apartment for decades. Then the white people moved in and police started showing up.* Buzzfeed News. https://www.buzzfeednews.com/article/lamvo/gentrification-complaints-\311-new-york.

Wainwright, O. (2017, Oct. 26). *'Everything is gentrification now': but Richard Florida isn't sorry.* The Guardian. https://www.theguardian.com/cities/2017/oct/26/gentrification-richard-florida-interview-creative-class-new-urban-crisis.

Wojcik, S. & Hughes, A (24 April 2019). *Sizing up Twitter users.* Pew Research. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

Zuk, M., Bierbaum, A. H., Chapple, K., Gorska, K., & Loukaitou-Sideris, A. (2017). Gentrification, displacement, and the role of public investment. *Journal of Planning Literature*, 33(1), 31–44. https://doi.org/10.1177/0885412217716439.

9 Appendix

9.1 Full List of Features for Topic Models

Note: Several of the topic distribution features included in the model are indicated with . . . in the model due to lack of space.

	Model Type		
	Baseline	Base & Non-Twitter	All Features
Feature	Median Rent	Median Rent	Median Rent
	Percent Black	Percent Black	Percent Black
	Percent White	Percent White	Percent White
	Percent University-Educated	Percent University-Educated	Percent University-Educated
	Percent Unemployed	Percent Unemployed	Percent Unemployed
	Percent of Households that are Renting	Percent of Households that are Renting	Percent of HHs Renting
		Trend in Comm Addresses	Trend in Comm Addresses
		Trend in Res Addresses	Trend in Res Addresses
		ST Vacant Res Trend	ST Vacant Res Trend
		ST Vacant Comm Trend	ST Vacant Comm Trend
		LT Vacant Res Trend	LT Vacant Res Trend
		LT Vacant Comm Trend	LT Vacant Comm Trend
		Crime Trendline 06-10	Crime Trendline 06-10
		Crime Trend 11-13	Crime Trend 11-13
		Felony Trend 06-10	Felony Trend 06-10
		Felony Trend 11-13	Felony Trend 11-13
		Crime per capita (PC) 06-10	Crime per capita (PC) 06-10
		Crime PC 11-13	Crime PC 11-13
		Felonies PC 06-10	Felonies PC 06-10
		Felonies PC 11-13	Felonies PC 11-13
		Noise Reports Trend 11-13	Noise Reports Trend 11-13
		FHV Complaints 11-13	FHV Complaints 11-13
		Bike Rack Complaints 11-13	Bike Rack Complaints 11-13
			Eng-Lang Tweet Share Trend
			Total Tweet Count Trend
			Tweet Count-Beginning
			Tweet Count-End
			Average Tweet Count
			Topic 0 - January 2011...
			Topic 1 - January 2011...
			Topic 44 - December 2013

9.2 Correlation Heatmaps for Mean and Max Topic Distributions

9.2.1 Mean Topic Distribution Correlation Heatmaps

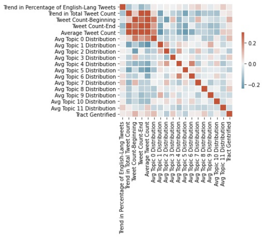


Figure 30: *Mean Part 1*

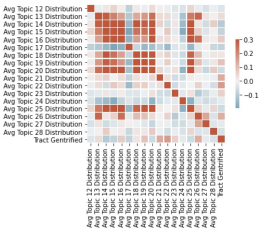


Figure 31: *Mean Part 2*

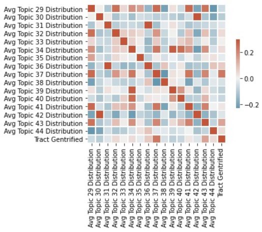


Figure 32: *Mean Part 3*

9.2.2 Max Topic Distribution Correlation Heatmaps

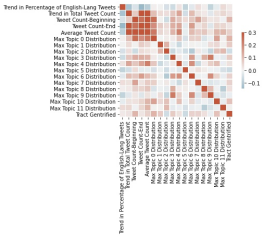


Figure 33: *Max Part 1*

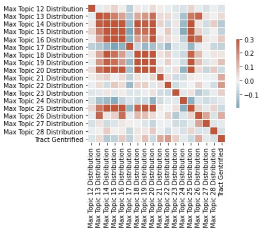


Figure 34: *Max Part 2*

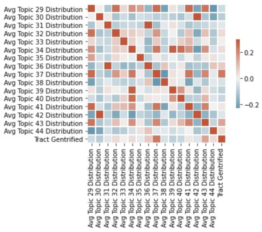


Figure 35: *Max Part 3*

9.3 Extra Model Results on the Test Set

Table 8: Comparing Results of Models

	Baseline	Twitter	Twitter Only	Twitter Only Upsampled
Logistic Regression	0.55	0.6736	0.6722	0.6806
Random Forest	0.4868	0.7257	0.7403	0.6299

9.3.1 Baseline Model Confusion Matrix: No Upsampling

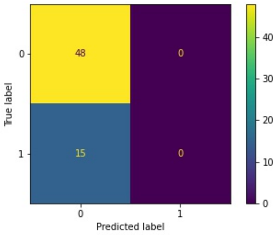


Figure 36: *Logistic Regression Baseline*

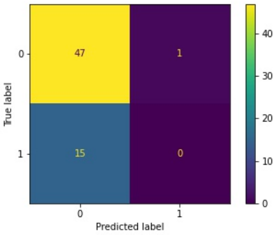


Figure 37: *Random Forest Baseline*

9.3.2 Full Model Confusion Matrix: No Upsampling

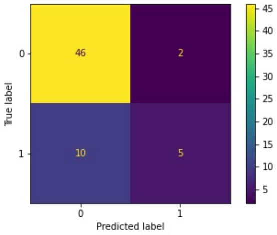


Figure 38: *Logistic Regression Twiiter*

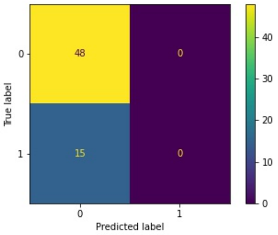


Figure 39: *Random Forest Twitter*

9.3.3 Twitter Only Model Confusion Matrix: No Upsampling

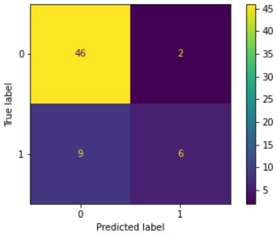


Figure 40: *Logistic Regression Twitter Only*

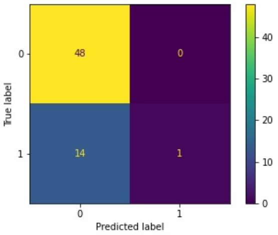


Figure 41: *Random Forest Twitter Only*

9.3.4 Twitter Only Model Confusion Matrix: Upsampling

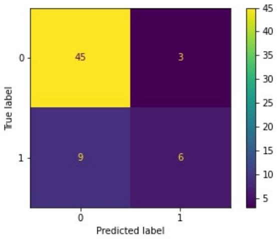


Figure 42: *LR Twitter Only Upsampled*

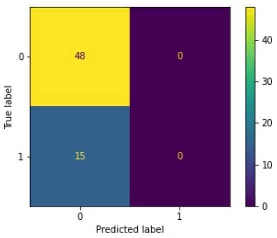


Figure 43: *RF Twitter Only Upsampled*

9.4 Feature Importance Values

Note that N/A does not mean the feature importance was not computed; rather, N/A indicates a feature whose importance is in the top 15 for one model but not the other one. For example, Percent Black is the 15th most important feature in the full Twitter LR model, but it is not in the top 15 for the RF counterpart.

Table 9: Selected Feature Importance for Baseline and Non-Twitter Models

Top Features	LR Beta Value	RF Gini Importance
Baseline Model		
Percent Black	0.47262	0.19535
Percent White	0.211543	0.179375
Percent of HHs Renting	0.211147	0.156159
Percent Bachelors	-0.00880933	0.14629
Median Rent	-0.0667857	0.150078
Percent Unemployed	-0.0930545	0.172748
Non-Twitter Model		
Felonies PC 06-10	2.1207	0.0406
Crime PC 06-10	-1.5589	0.0388
Noise Reports Trend 11-13	0.8709	0.0821
Felonies PC 11-13	-0.7184	N/A
Crime Trendline 06-10	0.53	N/A
Percent Black	0.5134	0.0825
Felony Trendline 06-10	-0.4712	0.0446
Percent Bachelors	-0.3057	0.0451
Crime Trendline 11-13	0.2908	0.0433
Percent of HHs Renting	0.2342	0.0409
Crime PC 11-13	-0.1803	0.0468
Percent White	0.157	0.0539
Bike Rack Complaints Trend 11-13	-0.1551	N/A
Trend in Residential Addresses	0.1189	0.0439
Percent Unemployed	-0.1132	0.0563
Median Rent	N/A	0.0496
Felony Trendline 11-13	N/A	0.0467
Trend in ST Vacant Residential Addresses	N/A	0.0413

Table 10: Selected Feature Importance for Full Twitter Models

Top Features	LR Beta Value	RF Gini Importance
Topic 3 Distribution: 2011-12	0.3105	N/A
Topic 39 Distribution: 2013-07	-0.276	N/A
Topic 3 Distribution: 2013-05	0.2431	0.0033
Noise Reports Trend 11-13	0.2364	0.0034
Topic 25 Distribution: 2012-04	0.2327	N/A
Topic 32 Distribution: 2012-07	-0.2111	N/A
Topic 17 Distribution: 2011-09	-0.2081	N/A
Topic 16 Distribution: 2013-01	0.207	N/A
Topic 17 Distribution: 2013-08	-0.2052	N/A
Topic 19 Distribution: 2012-03	0.2009	N/A
Topic 23 Distribution: 2013-10	0.2005	N/A
Percent Bachelors	-0.1988	N/A
Topic 38 Distribution: 2013-10	0.1983	N/A
Topic 22 Distribution: 2013-11	0.1957	N/A
Topic 4 Distribution: 2013-11	-0.1944	N/A
Topic 3 Distribution: 2012-07	N/A	0.0051
Topic 20 Distribution: 2013-07	N/A	0.0051
Topic 30 Distribution: 2012-05	N/A	0.0049
Topic 22 Distribution: 2013-12	N/A	0.0047
Topic 35 Distribution: 2013-07	N/A	0.0044
Topic 44 Distribution: 2013-11	N/A	0.004
Topic 30 Distribution: 2013-07	N/A	0.0037
Topic 39 Distribution: 2013-11	N/A	0.0037
Topic 2 Distribution: 2013-10	N/A	0.0032
Topic 6 Distribution: 2013-08	N/A	0.0031
Topic 22 Distribution: 2012-01	N/A	0.0031
Topic 22 Distribution: 2012-11	N/A	0.0031
Percent Black	N/A	0.0031

9.5 Full Topic Keyword List

Table 11: List of Topic Keywords: Topics 0-38

Topic Number	Keywords
0	parkside, rooster, virgo, wholesale, breezy, costco, arie, rosedale, panera, thevoice
1	contemporary, lic, jackson, roosevelt, hillside, esplanade, temporarily, feedback
1 (continued)	ranch, baychester
2	woodside, ditmar, pisce, briarwood, youuu, montauk, soundtracke, youuuu, ik, regent
3	beacon, apr, loew, nostrand, pourhouse, facility, athletic, checkout, pickup, kingdom
4	blockhead, sheepshead, canal, balloon, coal, veteran, ritz, cafeteria, lombardi, thrift
5	jobcircle, analyst, trndnl, java, sunnyside, misi, cybercoder, architect, guidance, technical
6	essex, roaster, ludlow, eventsinnewyork, dumple, splash, tropical, rockwood,
6 (continued)	stumptown, baking
7	highline, ll, gaga, wid, bbm, whine, twit, gyal, np, di
8	rego, wakefield, oo, goddess, worship, premier, qpr, penalty, manchester, liverpool
9	pond, severe, cathedral, translation, apocalypse, cnty, pod, thunderstorm, kip, ontheb
10	dumbo, mercede, bloomingdale, townhouse, sprinkle, runway, fashionweek
10 (continued)	interactive, blogger, creator
11	boardwalk, niall, louis, cyclone, mitt, roller, belieber, coney, mermaid, idol
12	tompkin, ellen, buddha, gyro, halal, stardust, frite, eastvillage, alpha, signature
13	waverly, stonewall, astor, papaya, peel, grove, rebel, ippudo, westvillage, niteline
14	moma, rockefeller, observation, highline, greenhouse, rink, promenade, mckittrick
14 (continued)	lego, skating
15	bedford, trinity, stuyvesant, cobble, smorgasburg, homage, van, marble, gutter, grove
16	greenmarket, handmade, caroline, tkts, assembly, sandal, specifically, tapa
16 (continued)	unionsquare, triangle
17	teamfollowback, followback, wht, billiard, cnt, nf, williamsbridge, pelham, bak, thnx
18	hospitality, blink, westway, keen, swift, stronglove, skating, atrium, birdland, smw
19	parson, narrow, knit, mercer, riverdale, toll, ribbon, roeble, villagevine, spuyten
20	flatiron, kindle, ebook, tube, dekalb, barbecue, auditorium, nightclub, ctr, carnegie
21	giveaway, doughnut, newsie, clearview, mnhttn, sanctuary, nj, plz, lighthouse, phuket
22	seaport, knot, beekman, obamacare, relation, economic, journalism, voter,
22 (continued)	provider, conservative
23	getalljobs, tinyurl, accounting, sporting, equity, ecommerce, nursing, analyst
23 (continued)	healthcare, employment
24	rochdale, platinum, glee, hanson, pleasee, shiny, sew, martial, apparel, consultation
25	biergarten, meatpacke, gansevoort, meatpacking, snowpocalypse, hum, brass
25 (continued)	smorgasburg, vicious, cameo
26	wwe, moving, boxer, canvas, sp, richie, saddle, wrestling, faggot, swedish
27	fdr, lool, stroke, wagner, yogurt, xd, scar, exotic, loool, ahaha
28	resort, lavo, invisible, forbid, niketown, brenner, schermerhorn, regal, strand, revolve
29	gramercy, macy, rapture, columbus, pavilion, baruch, dinge, interaction, fabric, wafel
30	mph, humidity, temperature, meadow, expwy, manor, vanderbilt, steady, springfield, utopia
31	stern, sotu, pony, bruckner, ultra, recreation, spotlight, doo, alexander, jerome
32	timessquare, regal, imax, bubba, gump, tussaud, potter, avenger, foxwood, rpx
33	momofuku, banksy, artichoke, basille, gluten, pale, croissant, vscocam, aged, stout
34	jingle, jingleball, leffert, santacon, endomondo, phish, vsfashionshow, voicesave, bell, anarchy
35	northbnd, bnd, rv, nycha, armory, bypass, fabipop, satisfied, canal, whitehall
36	deegan, certificate, lololol, reform, getaway, shxt, mondrian, ctfu, ay, calf
37	glassland, goodman, maison, bergdorf, supernatural, landmark, experimental
37 (continued)	freeman, bi, audio
38	tide, concourse, rod, gowanus, redsox, oriole, dodger, loll, sanchez, nyy

Table 12: List of Topic Keywords: Topics 39-44

Topic Number	Keywords
39	arena, knickstape, kew, bally, pacer, isle, bruin, okc, islander, clipper
40	fordham, ue, circus, sail, parkchester, motorcycle, kingsbridge, fairway, quot, ram
41	cloister, sterling, chamber, earring, inwood, index, leisure, additional, vracework, unicorn
42	assault, standtoendrape, gowanus, sexually, roundup, thief, nypd, jury, santo, firefighter
43	harbor, dylan, serendipity, willy, arch, eventsinnewyork, strive, wicke, gateway, wagon
44	junction, thot, trill, foh, kendrick, disrespectful, bae, idgaf, nvm, matchless