Matching Homes and Homebuyers using Machine Learning
August 2018

1. Introduction

1.1 Background & Problem

Finding the right the right home is not a challenge to be taken lightly, whenever one is new to the experience or not. In a city as diverse and cosmopolitan as London, the variety of choices makes this endeavor even more arduous. Of course, affordability is a significant that dictates whether someone will be able to purchase or lease a home or not. However, another equally important factor--one that is often overlooked--is suitability.

Some measures of suitability includes:

- Facilities: shops, restaurants, parks/recreation facilities, cinemas, parking, etc
- Space: when buying a home for the long run, buyers look for spacious settings where they can possibly grow families
- Noise: pubs, clubs, football stadiums
- Transport: distance to tube or rail station

This project seeks to help home buyers find properties that match both desired price and neighbourhood features.

1.2 Objectives

The goal of this project is to help real estate agents better connect clients to properties based on features that they are looking for using machine learning techniques. .

2. Data acquisition and cleaning

2.1 Data Sources

Two datasets with the relevant were combined for the analysis. The first (D1) was sourced directly from HM land registry and the relevant csv.file was downloaded. This publicly accessible data contains information that was collected as part of the land registration process.[1] The second dataset (D2) was collected from doogal.co.uk which contained all postcodes in London.[2]

2.2 Datasets

[1] https://www.gov.uk/government/collections/price-paid-data
[2] https://www.doogal.co.uk/london_postcodes.php

Df_data_1: The full csv file downloaded from HM land registry included 2018 sale prices for all of Wales and England. The file was opened in excel, sorted, and a new file created with only information from Greater London. From the dataframe, irrelevant/redundant variables were also removed. These include PPD (Price Paid transaction) type, Record Status, POAN-Primary Addressable Object Name-typically the house number or name, and SOAN-Secondary Addressable Object Name where a property has been divided into separate units (for example, flats). For each transaction, the remaining information was provided.

- ID (Transaction ID)
- Date (Date processed: Day/Month/Year)

-Transaction Price
-Property classification (Property Type, Old/New Building, Tenure)
-Address information (Postcode, Street, Town/City, District, County)

Df_data_2: The full csv file for this dataset was also initially opened in excel for exploration. It contained all postcodes in Greater London both active and inactive. All inactive codes were removed. From the dataframe, irrelevant/redundant variables were also removed. This included ward, region, altitude, easting, northing, constituency, etc. For each postcode, the following information was left in the dataset.
-Postcode
-Latitude
-Longitude

Df_data_3: Cleaning/preparing the data for analysis in jupyter notebook
Df_data_1 and Df_data_2 were imported imported into jupyter notebook and joined via postcodes to produce df_data_3. Df_data_3 was then checked for postcode duplicates of which there were 52,034. This was as expected given that a single building with multiple flats/living areas had the same postcode. To rectify this, only one recording of sales price is used per postcode and the other duplicates dropped.

3. Methodology

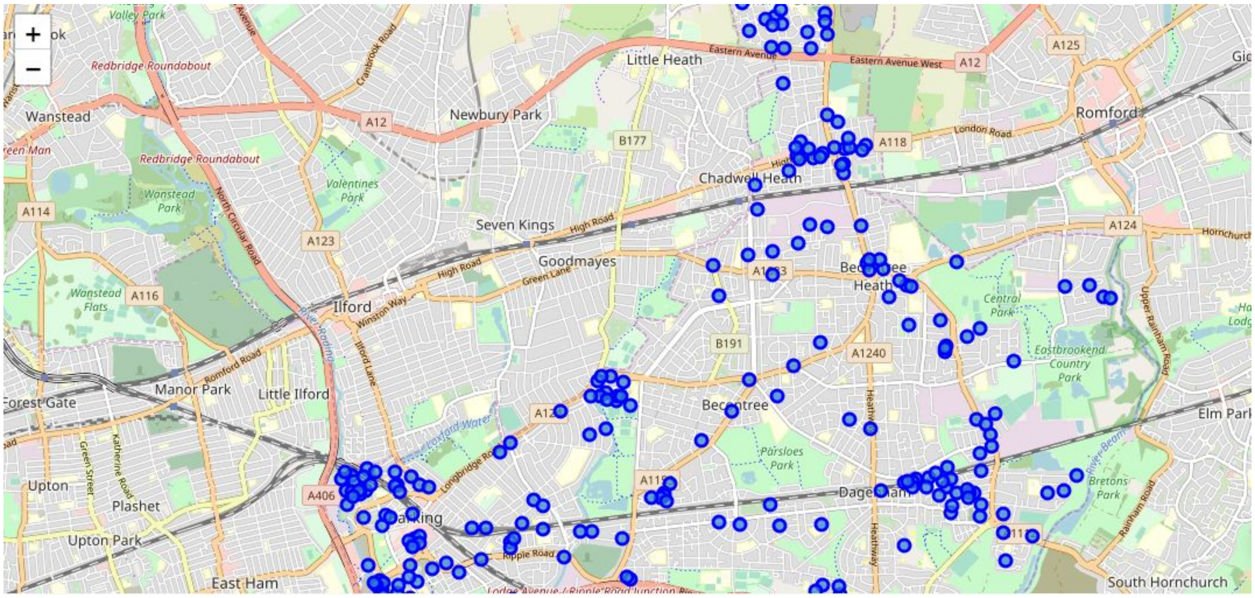
Exploratory Data Analysis

After data cleaning, there were 56385 rows/postcodes and 13 columns/features. Using the method value_counts, the numbers for the tenure status, property type and property age were measured. For the sake of this paper, we will be using a hypothetical case to demonstrate the machine learning algorithm to be used for the mobile application. A client is interested in finding a flat in London to lease under with sale value of under 400,000 pounds. A new dataset was created to include only the flats.

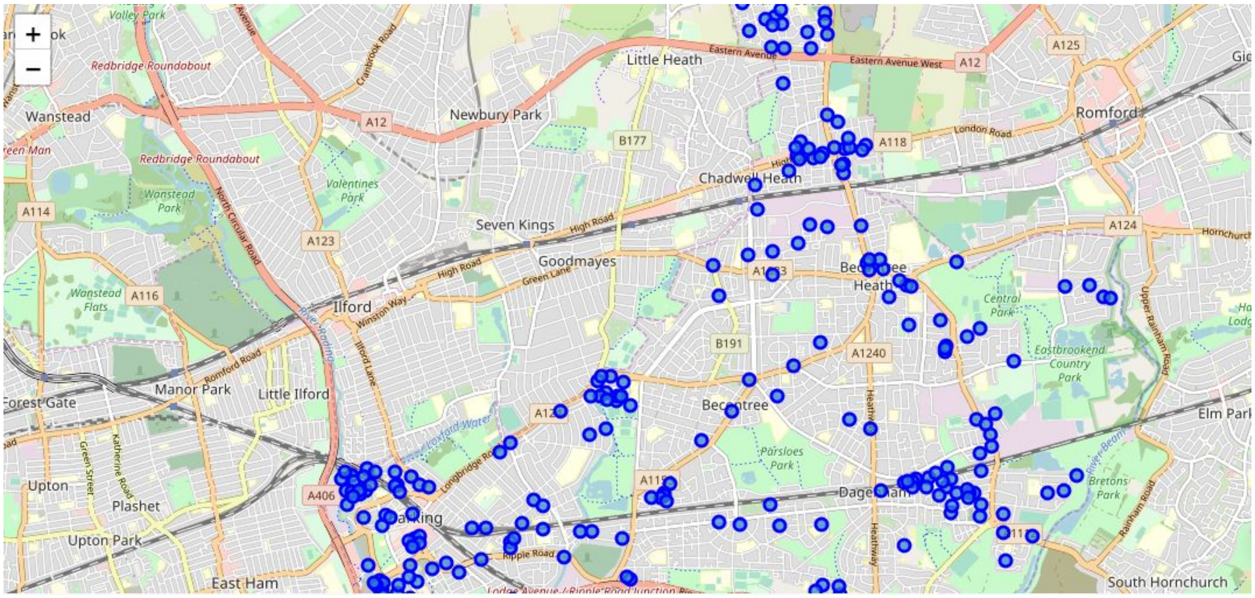
Table 1: Summary of Housing Top Ten Cheapest Districts Across London

District	Price	Latitude	Longitude
BARKING AND DAGENHAM	215287.3936	51.546836	0.120551
BEXLEY	232622.1975	51.459085	0.137659
HAVERING	235475.6915	51.574386	0.202485
CROYDON	260489.1525	51.377398	-0.090463
GREENWICH	261457.0153	51.476143	0.049714
ENFIELD	267108.1709	51.643028	-0.077646
HILLINGDON	267435.8518	51.539299	-0.440874
SUTTON	269644.5216	51.364263	-0.181839
REDBRIDGE	270996.4715	51.582171	0.065969
HOUNSLOW	272027.224	51.468124	-0.360679

Map Showing Flat Locations Across Barking and Dagenham



Map Showing Flat Location Across Bexley



The two areas with the cheapest flats are Barking and Dagenham, and Bexley. A function was created to get the top 100 venues that were within 500 metres radius of each postcode within Barking and Dagenham and the number of unique categories from all the returned venues determined. One hot coding was used to analyze the postcode data. The top ten venues for each postcode was printed and assessed.

Machine Learning Technique

Cluster analysis or *clustering* was used to compare the two areas. This method is appropriate because we are attempting to find the spatial patterns in the data. K-means is the clustering algorithm that was used to divide postcodes into 10 classes/groups. There were two postcodes from the original Barking and Dagenham with no venue information from foursquared so these were dropped. The final 10 clusters were then visualized in a map. **All steps were repeated for Bexley. (**N.B. Foursquare was missing venue data for 11 postcodes in the original Bexley dataset and so these postcodes were also dropped).

Results

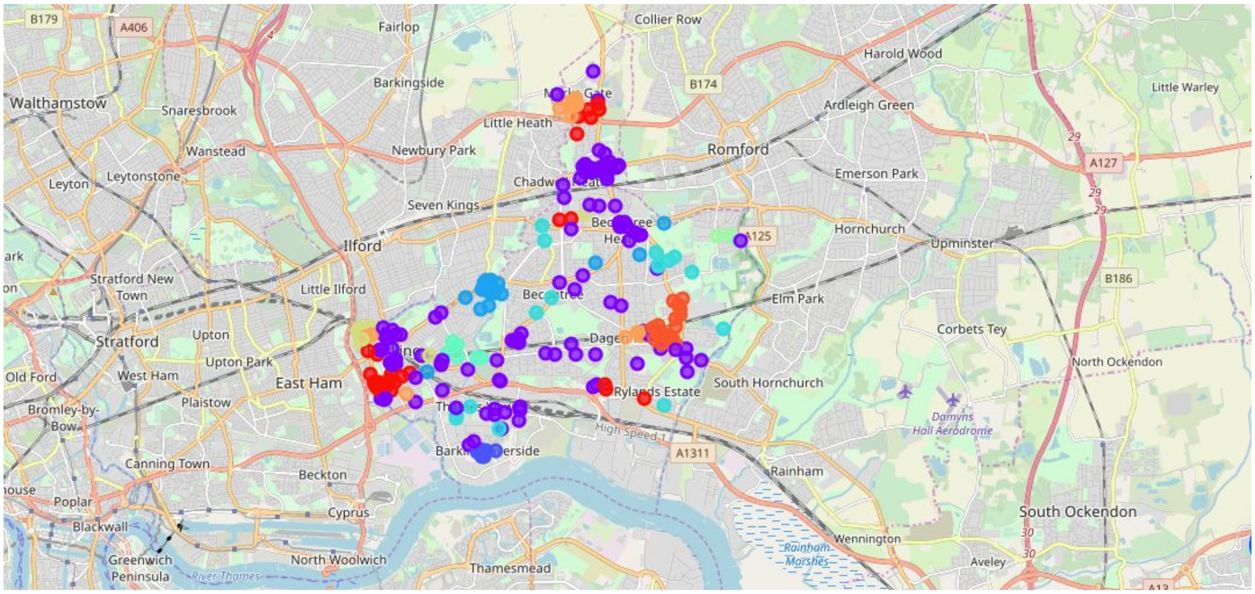
Barking and Dagenham

Clusters 1 and 3 also have disadvantages as the former has a high number of hotels which is a likely indicator of high traffic while the presence of many warehouses tends to indicate that the area is less residential. In terms of similarity, Cluster number 4 and 6,7, appear to be quite similar given that they same most popular type of venue. If having recreation areas/open areas nearby is important, venue five is convenient however the second most popular site in this cluster include construction and landscaping which indicates that this area might suffer from noise pollution. Though flats in cluster ten are positioned close to transport platforms, they bars are also heavily present in this area. Depending on the person, this might either be a pro or con. Clusters 2, 4, 6 and 7 are the most likely to appeal to most clients given that they have venues that are generally visited often. But choosing the right venue for a particular client will need to be guided by their own desires.

Table 2: Top 1st and 2nd Most Common Venue Category Per Cluster in Barking and Dagenham

Cluster	First Most Common Venue	Second Most Common Venue
1	Hotel	Indian Restaurant
2	Grocery Store	Pharmacy
3	Warehouse	Flea Market
4	Cafe	Park
5	Park	Construction and Landscaping
6	Cafe	Fish and Chips Shop
7	Cafe	Dry Cleaner
8	Hardware Store	Grocery Store
9	Grocery Store	Grocery Store
10	Platform	Bar

Map Showing Barking and Dagenham Clusters



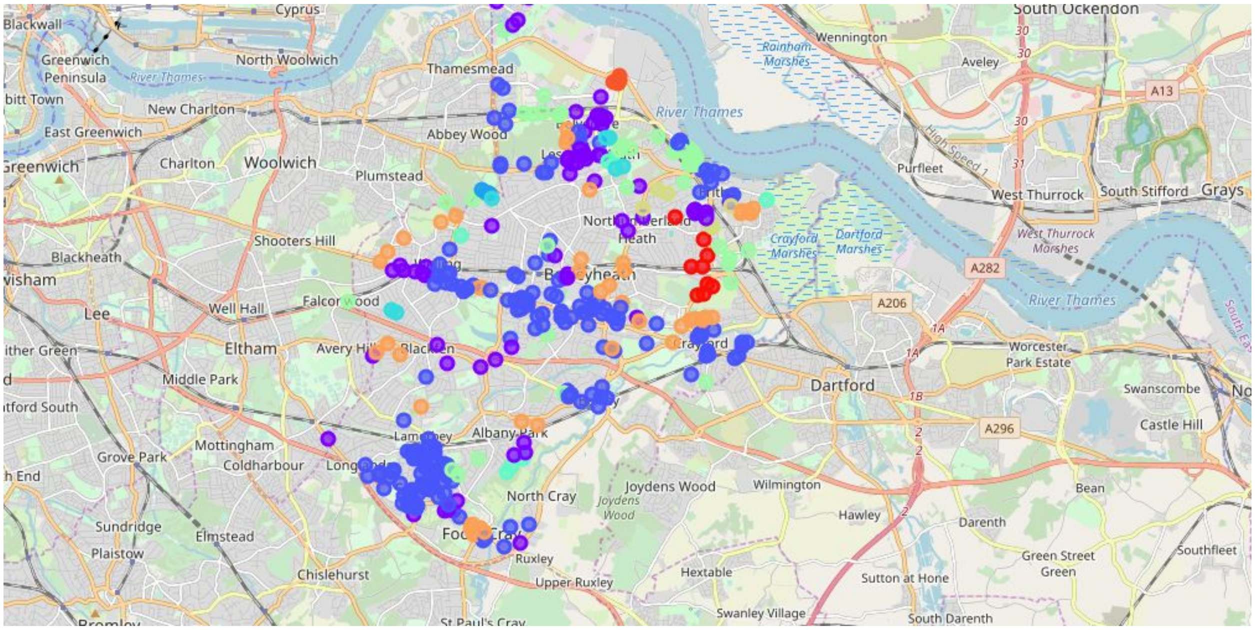
Bexley

Though cluster 1, 3 and 9 offer accessibility to food (Fish and Chips, Coffee Shop and Pub), the top two venues in each of these areas are the same suggesting little food variety which might be a limitation for some people. In contrast, clusters 2 and 10 both have food venues as their top two most popular spots, but both have different venues in 1st and 2nd place thus offering the food variety often sought by homebuyers. Though both cluster 4 and 6 have zoe exhibits as their second most common venues, the most popular venues these areas make them unattractive. For cluster 4, the most popular venue in is a “forest” suggesting that this area is in a more secluded region, possibly more difficult to access while the most popular venue in cluster 6 is some form construction/landscape suggesting a high noise area. Both cluster 5 and eight offer the entertainment amenities of 4 and 6 with the benefit of having food (a restaurant and grocery respectively) close by making themmore suitable as locations for a homebuyer.

Table 3: Top 1st and 2nd Most Common Venue Category Per Cluster in Bexley

Cluster	First Most Common Venue	Second Most Common Venue
1	Fish and Chips Shop	Fish and Chips Shop
2	Grocery Store	Fast Food Restaurant
3	Coffee Shop	Coffee Shop
4	Forest	Zoe Exhibit
5	Indian Restaurant	Zoe Exhibit
6	Construction and Landscaping	Zoe Exhibit
7	Pizza Place	Park
8	Grocery Store	Zoe Exhibit
9	Pub	Pub
10	English Restaurant	Coffee Shop

Map Showing Bexley Clusters



Discussion

In general, Bexley appears to have more postcodes/flats that are in areas with accessible food places (grocery stores, cafes, pubs, restaurants, etc). Moreover, Bexley seems to have more clusters that would generally be attractive to a home buyer (As seen in Table 2 and 3), though this advantage is only marginal. On the other hand, in Barking and Dagenham, there are more recreational areas such as parks and playgrounds. In addition, the clusters in Barking and Dagenham seem to offer a more variety in the popular venues so there are more options for clients who are looking for more than one specific feature in the neighbourhood that they want to live. Finally, as shown in Table 1, flat prices are on average lower in Barking and Dagenham compared to Brexley, and this could be a major selling point for undecided home buyers looking for cheap flat to rent in London. The recommendation for a homebuyer looking for cheap flat in London would therefore be to explore highlight clusters in Barkely and Dagenham (Table 2) to find appropriate housing.

Conclusion

Going forward, this type of research can be improved by combined the Clustering Machine Learning used in the project with other machine learning techniques to maximize solutions the “affordability and suitability” dilimea. At present the model relies on current housing prices to initially sort districts based on price, but it could be more useful for real estate agents to develop an algorithm to predict future prices in conjunction with the clustering method shown here to then connect the right buyer to the right property.