Opening a New Pub in London

Michael Turek

June 9, 2020

Introduction & problem

Pubs are an important part of culture and social life in London – in total, there are 4,500 pubs within the boundaries of Greater London and the density goes as high as 219 pubs in a single square mile in the City of London[1], suggesting that each pub is about a minute's walk from the next one. This number of pubs, however, is not distributed uniformly throughout the city – for instance, the further from the city center one goes, the lower the density of pubs they would see.

In general, we'd expect the number of pubs in a given neighborhood to be a function of many different variables that cover local demographics, level of disposable income, resident preferences, and other variables that specify how much business would be for a pub. Machine learning will allow us to analyze the neighborhoods of London at granular detail to try to understand these factors in more detail.

In addition, we can turn this understanding into a prediction of how many pubs a given neighborhood can support – if there are more pubs than we'd expect, they might be fighting for business and struggling to survive, and if there are fewer, we might be able to find an underserved market for a new pub. As the city wakes up from lockdown, understanding where (and if) to open a new pub will be valuable for hospitality entrepreneurs as well as existing pub operators.

Data acquisition & cleaning

To execute the analysis described above, we need to combine 3 separate datasets containing, 1) information about venues (and specifically pubs), 2) information about London neighborhoods, and 3) geographic information tying the other 2 datasets together.

[1] https://www.beerguild.co.uk/news/revealed-britains-pub-capitals-takes-top-spot/

The geographic data used was a static dataset of topoJSON boundaries[2] exported at the "Ward" level. With ~600 wards within Greater London, this level of detail allows us to carry out a more granular analysis than if we just focused on comparing London's 32 boroughs. On the other hand, we'll be less likely to encounter corner cases with the wards' average population of ~13,000 than if we ran the analysis at the level of LSOAs with an average population of ~2,000 people. To translate the map boundaries to GPS coordinates, we find the centroid of each ward's rectangular bounding box.

Having calculated the GPS centroid of each ward and its approximate radius, we can use the Foursquare API to download the set of venues in each ward. We use multiple calls to the API to bypass Foursquare limitations and download every single listed venue even in wards with more than 50 venues. We then clean the data and process it to identify the number of pubs in each ward based on Foursquare's category tagging

Finally, we use the ward atlas provided by the official London Datastore[3] to build a dataset of potential predictors. Out of the 66 columns provided (e.g., Population, % working age, average number of cars owned, etc.) we are able to retain 40 after cleaning and removal of multicollinear columns.

With the 3 datasets collected and cleaned, we can create a master dataset covering 40 predictors for each ward and a target variable which is the number of pubs in that ward. We then split our dataset into a training set with 499 samples and a test set with 125 samples.

[2] https://martinjc.github.io/UK-GeoJSON/
[3] https://data.london.gov.uk