Opening a New Pub in London

Michael Turek

June 13, 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.

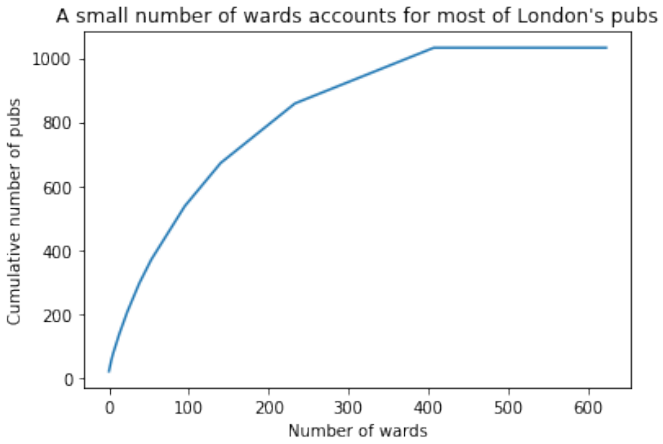
Methodology

As a first step, we can look at basic descriptive analytics to understand the geographic distribution of pubs in London. Even though some news articles mention there are up to 4,500 pubs in London, our analysis will be based on the 1032 pubs returned from the Foursquare API using the query above. Even though there are 600+ wards in London, we find that over 200 of them do not have a single pub. In fact, the top 10 wards with most pubs alone account for

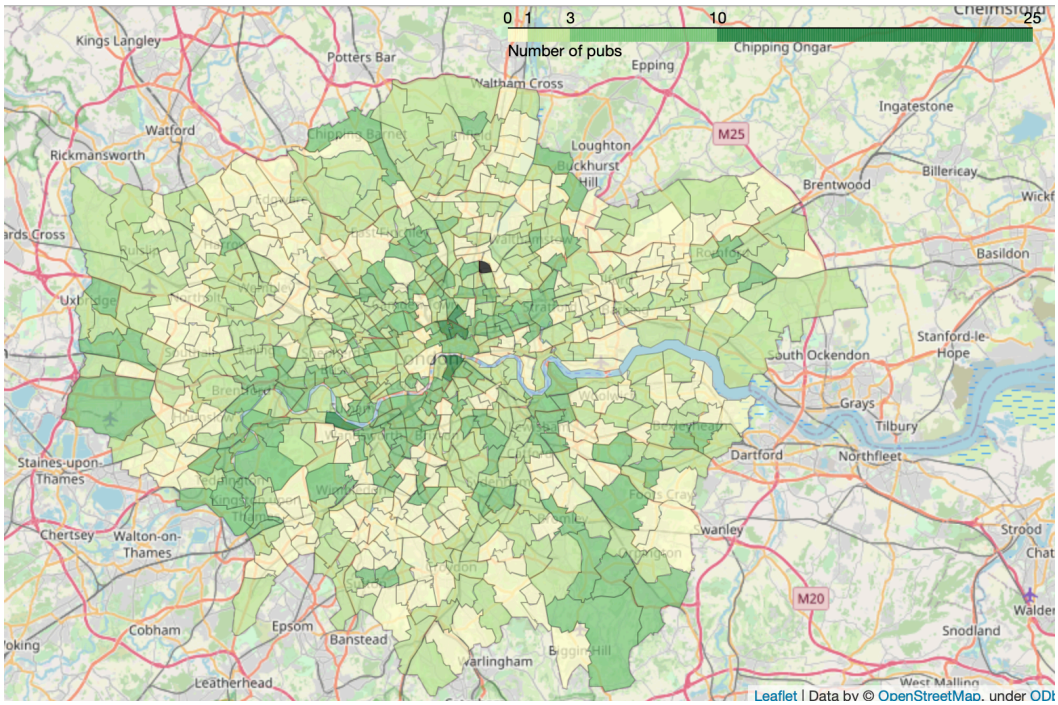
[2] https://martinjc.github.io/UK-GeoJSON/

[3] https://data.london.gov.uk

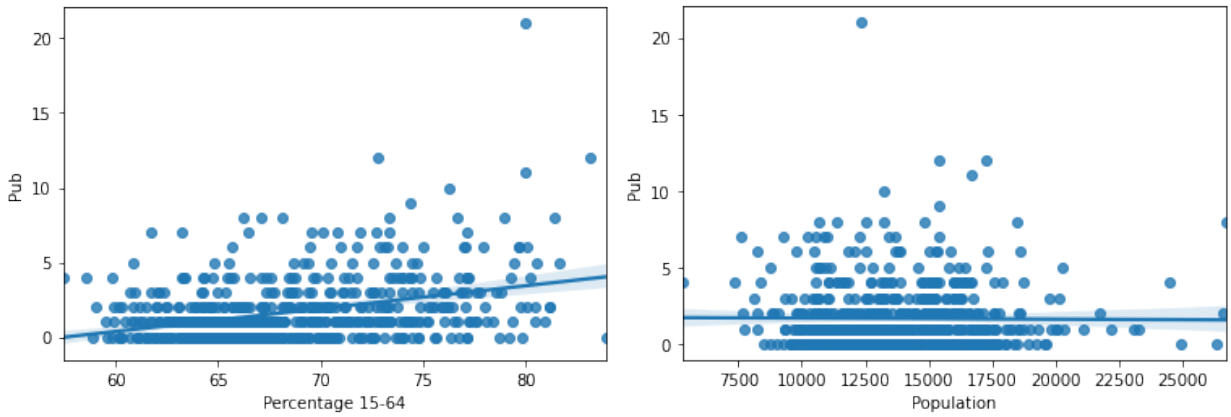
~10% of all pubs, and the top 100 wards account for more than half of all pubs. Clearly, there are large differences in the density of pubs among the different wards.



We can plot this distribution on the map to see if there are any geographic trends in the data. Overall, there seem to be a larger number of pubs in the center of London and along the riverfront. As we travel further out, the number of pubs generally decreases with some exceptions, such as the Heathrow airport at the western border of London. In addition, local centers such as Uxbridge, Chipping Barnet, Wimbledon, or Kingston upon Thames generally have a higher number of pubs than the suburbs that surround them.



Finally, we can look at correlations within our dataset to see if there are any specific characteristics that are strongly positively or negatively correlated with the number of pubs in a ward. We find that neighborhoods with high accessibility of public transport and a high share of working-age inhabitants tend to have more pubs on average. On the other hand, neighborhoods with many children under 15 and higher car ownership tend to have fewer pubs. While these trends could suggest a causal relationship – e.g., pubs clustering along railroad stations in the suburbs – they could also be a result a hidden variable where neighborhoods that are more "urban-like" have more pubs than those that are "suburban-like". Interestingly, there isn't a strong relationship between the size of population in a ward and the number of pubs, suggesting that a high number of people living in a given ward doesn't alone imply there is a market for pubs.

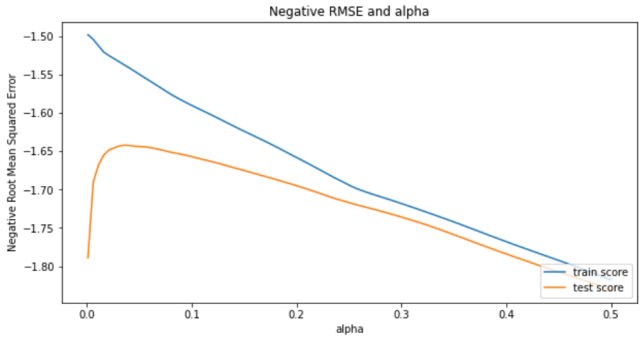


As the next step, we will use machine learning algorithms to try to predict the number of pubs in each ward using the full set of predictors we have available. We will start with the simplest model – Linear Regression – to set a baseline before moving on to the more involved techniques of LASSO Regression and Polynomial LASSO Regression. In the case of the LASSO regressions, we will use 10-fold cross-validation to set the regularization parameter which minimizes the out-of-sample root-mean-squared-error. Finally, once we select the best model, we will train it on the full dataset and calculate a prediction for each of London's wards. We will then compare the prediction with the actual number of pubs in the ward to identify whether the market is over- or under- saturated on a ward-by-ward level.

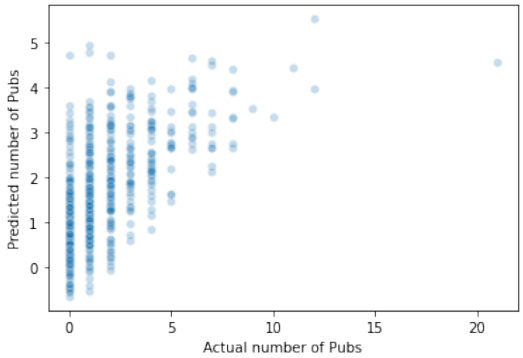
Results

Testing the different models, we find that all perform relatively similarly, giving us a stable prediction. The LASSO model with L1 regularization performs best out-of-sample with a shrinkage parameter or 0.036. While adding polynomial terms reduced the RMSE on the train set, it fails to improve the prediction on the test set, likely due to overfitting the features of our train data. Even a high level of regularization cannot offset the overfitting, resulting in a better performance for the base LASSO regression.

Model	OOS RMSE
Linear regression	2.31
LASSO regression	**2.28**
Polynomial LASSO regression	2.29

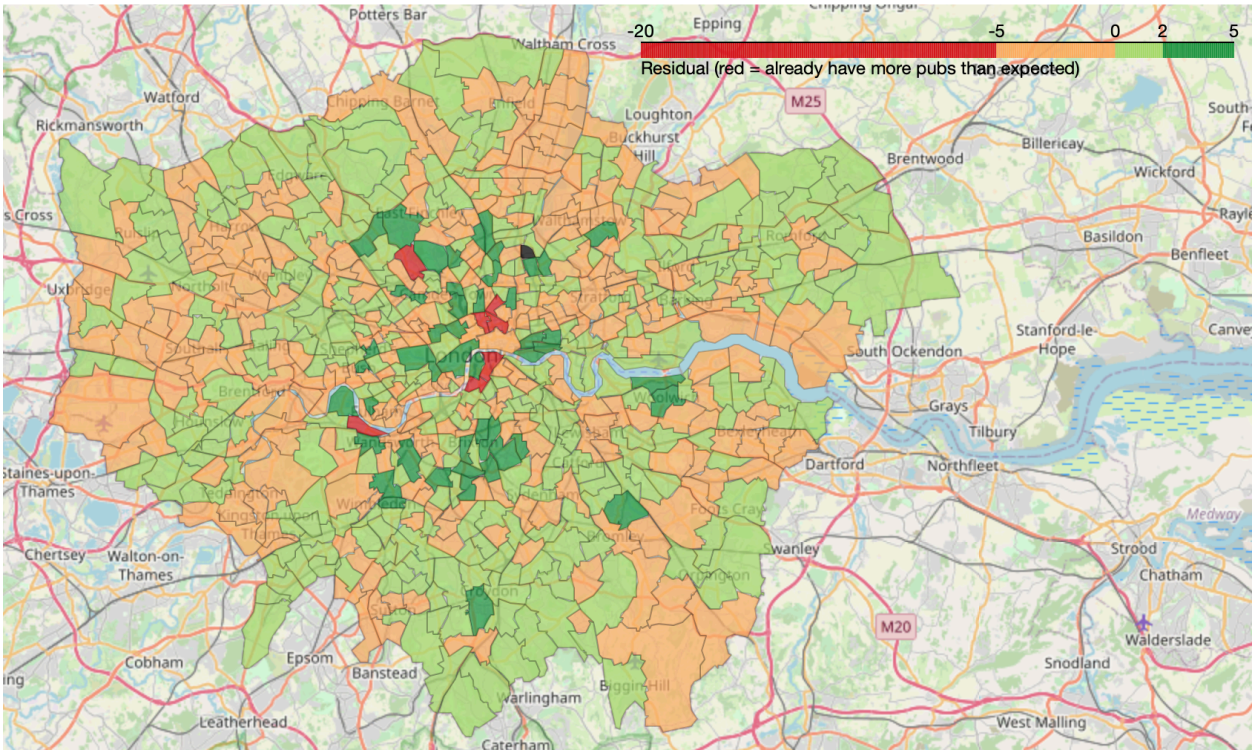


With this model, we can predict the number of pubs for each ward, resulting in the predictions shown below. For 342 wards, the model predicts a higher number of pubs than they already have, suggesting the areas might be underserved. On the other hand, there are 282 wards where there are currently more pubs than expected, suggesting they might be overserved.



Discussion

The analysis below gives us a good starting point to answer our original question – if we were to open a new pub in London, which location should we choose? The map below shows the geographic distribution of the regression residuals. Areas marked in red already have more pubs than predicted (and thus could be overserved) whereas areas marked in green show potential for the opening of additional locations.



Plotting the data on the map shows that the results of our analysis do not follow the same geographic trends as the distribution of pubs overall – namely, while some wards in central London (which overall had the most pubs) do have too many pubs, there are others where more could be opened and still have a market. For instance, even though the popular districts around Old Street and Shoreditch are starting to show oversaturation of pubs, the neighboring King's Cross area could accommodate more. The largest opportunity to open new pubs, however, seems to lie further out from the center, especially in South London where Brixton and its surroundings suggest large opportunity for new pubs. In the north, there is an opportunity for pubs to open in areas surrounding Hampstead Heath and compete for

customers with the oversaturated Hampstead Town. Finally, there is a pocket of opportunity in East London between the City of London and Canary Wharf which is currently underserved despite being positioned between the two financial centers.

Conclusion

In this report, we analyze the factors that affect the pub market in different wards in London. We use a combination of geographic and demographic data to train a machine learning algorithm which predicts whether each ward is overserved or underserved in order to determine the best location to open a new pub. On the borough level, we find that the most overserved borough is Islington while there is space for additional pub expansion in Croydon, Lambeth, and the Tower Hamlets.