Coursera Data Science Capstone – Islington Pubs

Business Case

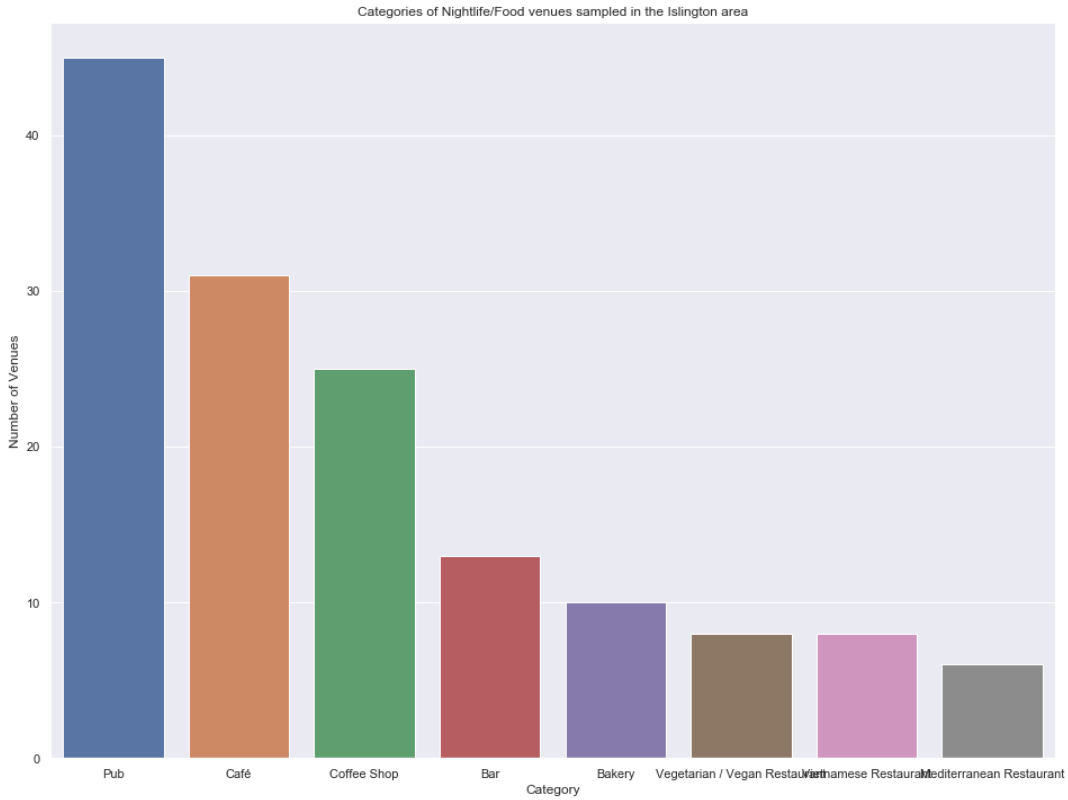
The business problem I am suggesting is that a friend wants to open a Pub in the Islington area of London. They want to know the most popular restaurant venue types already in the area (including competitive pubs) and also when different types of restaurants are popular so that they can target their marketing and resources accordingly.

Data Requirements

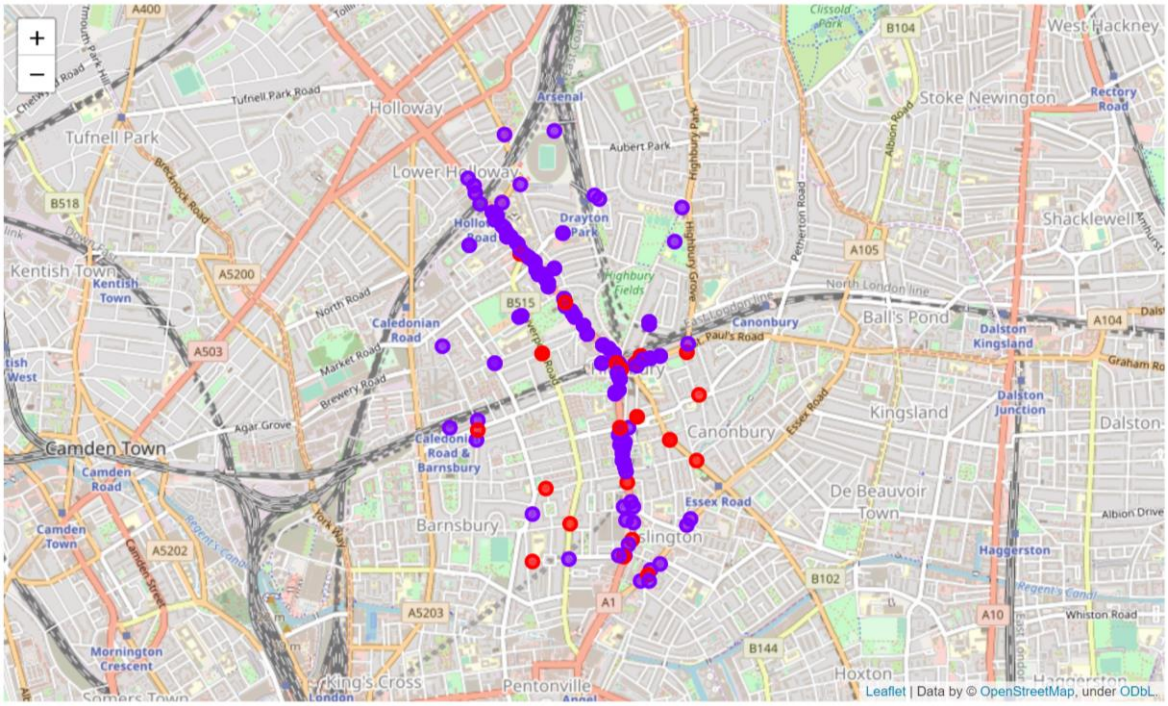
The data that I will need is a list of the restaurants and bars in the local area, the type of venue that they are, the times that they are open and the times that they are popular. I'll include data on other types of restaurants and bars so that my model learns about when food/drink venues are generally popular in the local area. Including data on other types of bars and restaurants will help to prevent my model from over-fitting. It may also highlight a gap in the market if there are other restaurants that are open and popular but existing pubs are not open.

Data Sourcing- Nearby Venues

Using the Foursquare API, I gathered data on 250 venues in the area of Islington (with a 1km radius) that were either nightlife or food venues, as these are the categories most similar to pubs. I noted that pubs where the largest category of venue within the sample:



I also visualised where the pubs were: it appears that they are evenly spread throughout the Islington area:

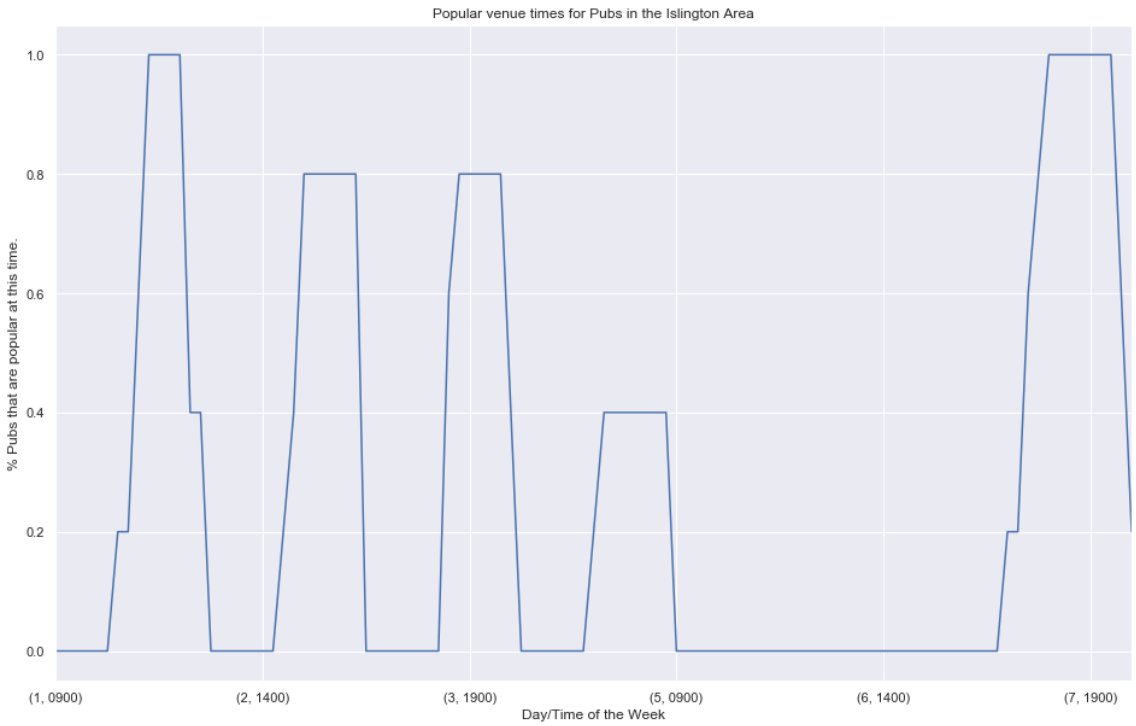


Data Sourcing – Venue Opening Times

Using the FourSquare 'hours' API endpoint, I pulled the opening hours and 'popular' hours for a sample of 38 venues in the area (due to API limits this was the maximum that I could access).

I wrangled this data considerably to identify all hours between 9AM and 11PM for each pub, and determined whether they were open and/or popular.

I examined the popular hours of the pubs to see whether it fluctuated throughout the week, producing a line plot of the proportion that were popular at any given hour, as below:



It appeared that the popularity of pubs definitely does fluctuate during the week, and therefore it was definitely worth examining the dates that pubs were likely to be popular, in order to avoid my friends overstaffing or wrongly marketing.

Machine Learning

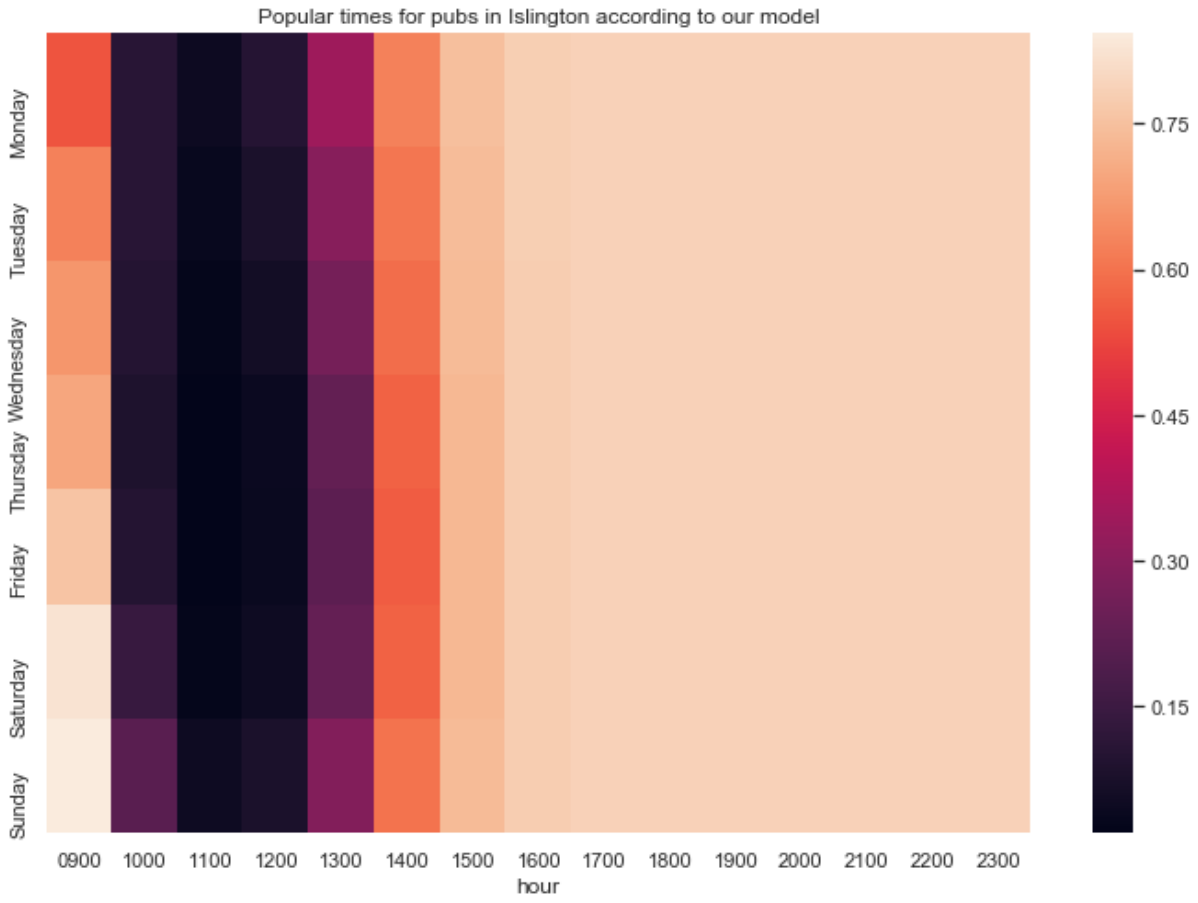
I built a 2.2k row dataset with 7 input features, and built several models to examine which modelled 80% of my dataset best.

The best model was a Support Vector Classifier with c=3.0. This achieved a 80.6% test f1-score, with 78.4% precision and 83.0% recall.

I then used the model to predict against 9AM-11PM every day of the week, to predict when a pub in the Islington area would be most popular. I pivoted this data to produce the following results:

hour	0900	1000	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000	2100	2200	2300
Monday	0.547583	0.108821	0.046148	0.100461	0.342903	0.626348	0.746624	0.779557	0.786769	0.787972	0.788122	0.788135	0.788136	0.788136	0.788136
Tuesday	0.626085	0.108206	0.036050	0.077020	0.301856	0.607638	0.741825	0.778585	0.786613	0.787954	0.788120	0.788135	0.788136	0.788136	0.788136
Wednesday	0.665456	0.098766	0.027213	0.058875	0.265782	0.589901	0.737406	0.777706	0.786473	0.787937	0.788118	0.788135	0.788136	0.788136	0.788136
Thursday	0.695773	0.085814	0.019742	0.043908	0.230987	0.571070	0.732797	0.776800	0.786330	0.787919	0.788117	0.788135	0.788136	0.788136	0.788136
Friday	0.758297	0.096209	0.018860	0.039651	0.217042	0.561992	0.730459	0.776327	0.786254	0.787910	0.788116	0.788135	0.788136	0.788136	0.788136
Saturday	0.824772	0.140326	0.026374	0.048477	0.234205	0.570475	0.732295	0.776657	0.786303	0.787916	0.788117	0.788135	0.788136	0.788136	0.788136
Sunday	0.842901	0.208802	0.048350	0.078519	0.290270	0.598603	0.739045	0.777965	0.786508	0.787940	0.788119	0.788135	0.788136	0.788136	0.788136

This produced the following heatmap:



Results

The clear conclusion is that pubs are generally more popular from 5PM onwards; and on weekends (particularly Sunday) throughout the day. While the model predicts that 9AM is also a popular time for pubs, this may be impractical to staff or promote as subsequent hours are likely to be unpopular.

Conclusion

This research gives clear, commercially valuable insights for my friends as to how to staff and market their pub, even before it has opened. That said, this is likely to be affected by seasonality (people may visit pubs for longer periods in winter) and there is a somewhat small sample size compared to the number of pubs in Islington. Consequently, further analysis would examine additional samples from the area as well as seasonal changes in popularity for pubs to give greater insight.