The Missing Irish Pub

IBM Data Science Professional Certificate Capstone Project

Christian Weymann

Executive Summary

We want to open a new Irish pub in New York City. For this enterprise to be a success, we have to find a good location. We will use Foursquare data to determine in what type of neighborhood (identified by the typical venues in this neighborhood) Irish pubs are currently established, to try and find a similar neighborhood with no Irish pub. We identify several of such opportunities, the most promising one being the Tribeca neighborhood in Manhattan according to our model.

1 Introduction

In a large city like New York, each neighborhood can have a different feel to it: some are purely residential, others have a lot of shopping opportunities, or business centers, or nightlife venues. Implanting a new business in the wrong type of neighborhood can be a fast road to failure, as customers will not think of going to that location for this type of service. It also drastically reduces the chances of being discovered by accident by a customer who was already in the area.

In this project, we will take the example of Irish pubs to demonstrate this principle. Using data from the Foursquare Places API, we will determine what kind of neighborhood would be a good location for a new Irish pub in New York City. To this end, we will try to predict the number of Irish pubs in each neighborhood based on the frequency of other types of venues present in that area. We can then compare the predicted number to the actual number to identify opportunities.

The same kind of analysis done here for Irish pubs could be done for any type of business. Any prospective business owner could benefit from reproducing it. It could also be used by local governments to try to get a more data backed understanding of the area under their responsibility, and what kind of businesses they need to attract if they want to change the feel of their neighborhood.

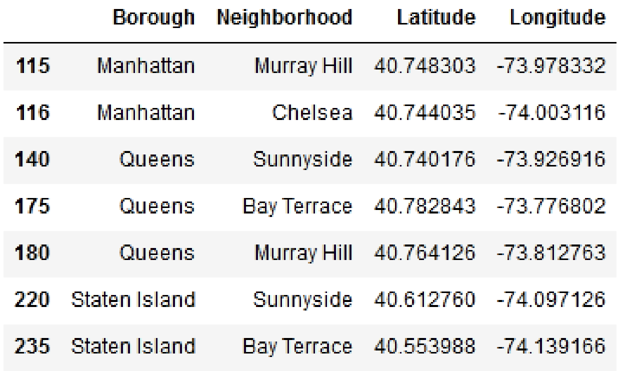
2 Data

2.1 Data Acquisition

Our data will mostly be loaded from the Foursquare Places API. Foursquare maintains a list of venues with a lot of rich data associated with each of them, such as ratings and comments from its users. However, we will be mostly focuse on core data about each venue: its geographical position, and its category. Foursquare lists the category of each venue both with a name and a unique category identifier which can be used in searches. We will take advantage of this to find all the Irish pubs in each neighborhood.

2.1.1 Neighborhood data

The geographical data of each neighborhood is taken from the json file used during the lab session of this course. It contains geographical information for the 306 neighborhoods of New York City. We extract the names of the neighborhoods and their boroughs, as well as the central latitude and longitude. These will serve as the positions around which to search for venues in our calls to the Foursquare API.



	Borough	Neighborhood	Latitude	Longitude
115	Manhattan	Murray Hill	40.748303	-73.978332
116	Manhattan	Chelsea	40.744035	-74.003116
140	Queens	Sunnyside	40.740176	-73.926916
175	Queens	Bay Terrace	40.782843	-73.776802
180	Queens	Murray Hill	40.764126	-73.812763
220	Staten Island	Sunnyside	40.612760	-74.097126
235	Staten Island	Bay Terrace	40.553988	-74.139166

Table 1: List of neighborhoods with the same name

For further analysis, we need a unique way to identify each of the neighborhoods. Using their names is not practical, as there are several pairs of neighborhoods with the same name in different boroughs, see Table 1. Instead, the index of this data frame will serve as a unique identifier for each neighborhood.

2.1.2 Pub data

We acquire the data for the pubs in each neighborhood using the "search" endpoint of the "venues" group of the Foursquare Places API. This endpoint allows us to pass a category ID as an argument to limit the results to venues of a particular category. We use this option to find all Irish pubs within 500 meters of the center of each neighborhood, and load them into a data frame for further analysis.

2.1.3 Typical venues data

To find the typical venues in each neighborhood, we use the "explore" endpoint of the "venues" group. This returns up to one hundred venues that are close to the center of each neighborhood, but are also considered relevant by Foursquare. Furthermore, this endpoint does not require to be passed an explicit radius to be searched, instead defaulting to a suggested radius based on the density of venues in the area. This allows us to get a better picture of what venues are actually typical in each neighborhood.

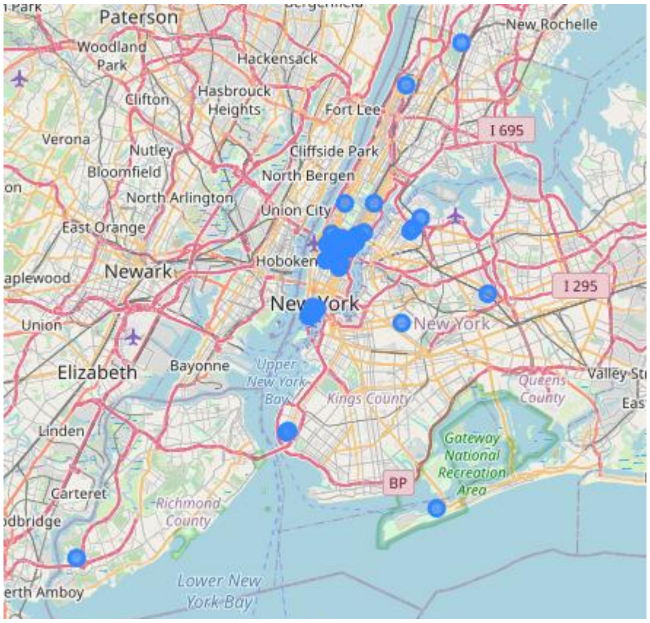


Figure 1: Locations of all the Irish pubs in New York City

2.2 Data Exploration

To get a first idea of our pub data, we plot the position of each pub on a map of New York City (Figure 1). We can see that the pubs are very inhomogeneously distributed, with most of them being located in central Manhattan.

Our target variable is going to be the number of Irish pubs in a given neighborhood. We can easily compute it by counting the entries in our pub data frame that are located in each neighborhood. The distribution of this target variable is shown in Figure 2. As we could already guess from the map, most neighborhoods do not have any Irish pubs, whereas a few neighborhoods have many. This is expected for a variable describing the number of occurrences of a rare event, which is best described using a Poisson distribution.

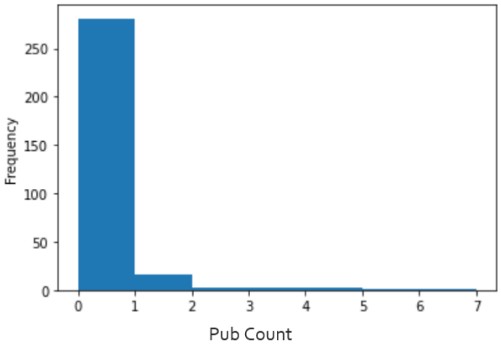


Figure 2: Histogram of the number of Irish Pubs in each neighborhood

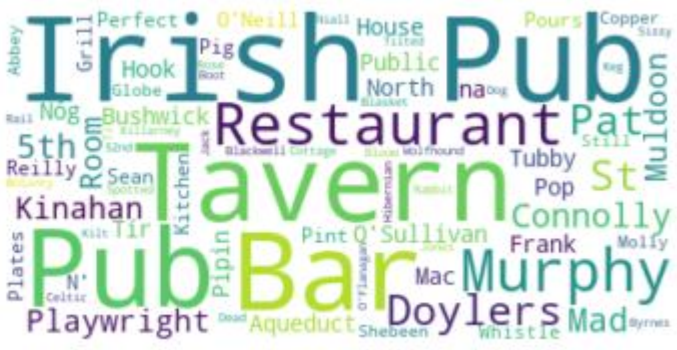


Figure 3: Word cloud generated from the names of the Irish pubs in New York City

Finally, let us search for inspiration to name our new pub. We generate a word cloud from the names of all the pubs in our data frame, shown on Figure 3. Unsurprisingly, the words “Pub”, “Bar”, and “Irish Pub” are very common. “Tavern” seems to also be frequently used, indicating that many of these venues go for an “old timey” feel. Other notable categories are Irish sounding names, and beer brand names. The word “pint” also makes an appearance.

2.3 Feature Engineering

2.3.1 Frequency encoding

To transform our list of venues and their categories into numerical data that can be used for machine learning, we extract the frequency of each category in each neighborhood. This is done using a two-step process: we first encode the category of each venue numerically, using one-hot encoding. In this scheme we create as many “dummy” attributes as there are different categories, and set them all to zero, except the one corresponding to the category of the venue under consideration. In the second step we average these dummy attributes for each neighborhood. In effect, this counts the number of venues of each category in each neighborhood, divided by the total number of venues in that neighborhood, yielding the frequency of that category.

2.3.2 Grouping similar venue categories

The categories used by Foursquare are quite detailed, leading to a large number of different categories and very sparse frequency lists for each neighborhood, with most categories not appearing at all in any given neighborhood. This could be detrimental to the performance of the machine learning algorithms.

Another issue is that we risk hiding the forest with the trees: the general trend of there being a lot of bars in a neighborhood might not get picked up if these bars are scattered over 17 different categories, see Table 2.

To solve this issue, we set out to group categories into more general "parent" categories. Fortunately, Foursquare groups its categories in a hierarchical structure, available through the "categories" endpoint of the "venues" group of the API. The response is a nested list of dictionaries, where each categories has attributes for its name and id, but also a "categories" attribute containing a list of subcategories, which may be empty.

Grouping the categories therefore comes down to finding the category of each venue in our list in this hierarchy, to determine all its parent categories, and then to consider a given depth in the category tree as the new category for our venue. Due to the fact that most categories in our dataset where either one of the 9 "basic" categories of Foursquare ("Nightlife Spot", "Outdoors & Recreation", "Residence", …) or a direct subcategory of one of these, we decided to only consider the groupings based on the first and second level of the hierarchy. Some examples of category groupings using the second level are given in Table 2.

Parent category	**Child categories**
Bar	Pub, Bar, Sports Bar, Beer Bar, Wine Bar, Beer Garden, Cocktail Bar, Speakeasy, Karaoke Bar, Hookah Bar, Whisky Bar, Tiki Bar, Dive Bar, Gay Bar, Hotel Bar, Sake Bar, Beach Bar
Movie Theater	Multiplex, Indie Movie Theater, Movie Theater
Stadium	Tennis Stadium, Baseball Stadium, Track Stadium, Basketball Stadium, Stadium

Table 2: Examples of grouped categories

3 Methods

3.1 Model selection

We want to train a model on the current data of neighborhoods in New York City to figure out where an Irish pub could be added, that is find a neighborhood similar to another one with more Irish pubs. We see three main types of models which could achieve this goal:

1. A classification model trained to differentiate neighborhoods with at least one Irish pub from neighborhoods with no Irish pubs. We could then see which neighborhood with no Irish pub would be classified as having one, indicating the model sees similarities with neighborhoods with an Irish pub. This option would not allow us to find an opportunity in a neighborhood that already contains an Irish pub. Furthermore, the training would be complicated by the extreme imbalance in the classes: out of 306 neighborhoods, only 25 have at least one Irish pub.
2. A recommender system, by treating the Irish pubs as a user, and the neighborhoods as products to be recommended to them. We can then just see what neighborhood the algorithm would recommend next. A starting point for a similarity metric between neighborhoods could be the

cosine similarity between their frequency vectors, which is implemented in scikit-learn. This approach might be biased towards neighborhoods which already have several Irish pubs.

3. A regressor model that fits the distribution of the number of pubs in each neighborhood. As we've seen above, the distribution seems to follow a Poisson distribution, which can be fitted using the PoissonRegressor from scikit-learn. We can then look at the difference between the predicted and the actual number of Irish pubs in each neighborhood to identify opportunities.

Due to the limited timeframe of this assignment, we decided to only implement option 3, as it seemed the most likely to succeed.

3.2 Hyperparameter optimization

Using this model, two parameters remain to be optimized: a regularization parameter alpha used by the PoissonRegressor class, and which grouping of the categories to use to get the optimal result. We therefore fit the regressor using three different datasets: one based on the original categories, one based on the first, and on based on the second level of the category hierarchy. In each case, the dataset is split into a training and a testing set. The optimal alpha parameter is then obtained by cross-validation on the training set, and the optimized regression is scored on the testing set using mean Poisson deviance. Since the data is heavily unbalanced, we compare the results of each of the regressors to the score of a dummy regressor that always guesses the mean value over the neighborhoods.

The dummy regressor obtains a mean Poisson deviance of 0.95 (between 0 and infinity, lower is better). The regressor fitted on the data obtained from grouping to the first level of the category hierarchy performs slightly worse than that, with a mean Poisson deviance of 1, however, the regressors fitted with the original categories and the second level grouping both outperform the dummy regressor with a score of 0.74 and 0.67 respectively. We therefore keep those two regressors for our final analysis, and refit them using the whole dataset and the optimized alpha parameter.

4 Results and Discussion

We can now compare the predicted number of Irish pubs for each neighborhood to the actual number, to create a metric of how highly each neighborhood is recommended by each model. When we compare the top ten recommendations for each model, some differences become apparent. While some neighborhoods make both top tens (namely Tribeca in Manhattan, and North and South Side in Brooklyn), the model trained on the grouped categories sees negative opportunity for the top neighborhood predicted by the model trained on the original categories. To better understand what is going on, let's compare the predicted distribution for each of these models with the actual distribution.

As we can see on Figure 4, while both models overestimate the number of neighborhoods with no pubs, the “Parent” model, fitted on the grouped categories, reproduces the distribution at a lower number of pubs more faithfully. However, it does not capture the existence of outlier with a very high number of pubs in one neighborhood. The “Child” model, which was fitted on the original categories, predicts outliers, but it overestimates the number of pubs in them, which is what leads it to recommend these neighborhoods. There seems to be an overfitting where a very specific type of venue is considered highly likely to be associated with Irish Pubs in the “Child” data, which is smoothed out in the aggregated “Parent” data. For our purposes, the “Parent” model therefore seems more reliable, which can also be seen by its slightly better performance during benchmarking.

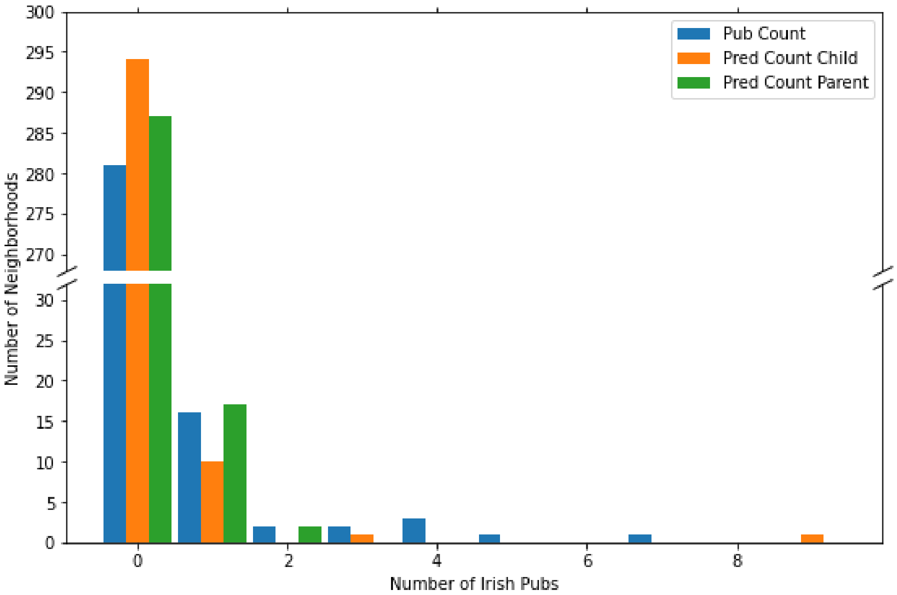


Figure 4: Actual distribution of Irish pubs (in blue), and predicted distribution from the model fitted with the original categories (in orange) and the grouped categories (in green).

Figure 5 shows the top ten locations for a new Irish pub, color coded from most recommended (orange) to least recommended (dark blue). We can see that the neighborhood of Tribeca in Manhattan is by far the most recommended spot for a new Irish pub, that is that our model thinks it is the neighborhood with no Irish pubs that is the most similar to neighborhoods with an Irish pub. For the purposes of this assignment, this is therefore our recommended location.

5 Conclusion and Perspectives

In this study, we modeled the distribution of Irish pubs in New York City using data on the types of venues present in each neighborhood. We were able to increase our models accuracy by grouping similar categories of venues together. These models could help an aspiring Irish pub owner by making a recommendation on where to open their business. Based on the other venues present there, we believe that Tribeca, in Manhattan, which does not currently have an Irish pub, should have at least one.

These results could certainly be refined by trying any of the other models presented in Section 3.1. We could also try to acquire more demographic data from public sources, such as median household income or land prices, to get a more accurate description of each neighborhood.

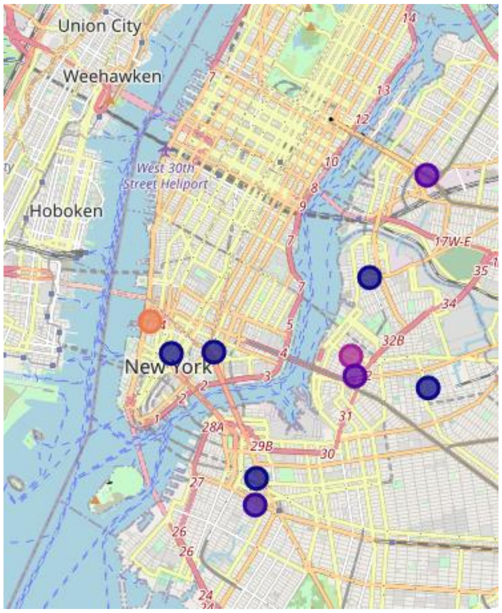


Figure 5: Map of the top ten most recommended locations for a new Irish pub. Orange is the highest and dark blue the lowest recommendation level.