Predicting Local Business Popularity during the Olympics through Geospatial Features

Shizhuo Duan, Vivek Gupta, Xu Han, Darshan Solanki

July 19, 2020

1 Problem Statement

The Olympic Games are viewed as a considerable investment in the host city due to the infrastructure spending and increase in tourism. Considerable research has been performed regarding the impact of hosting the games on the city as a whole, but less is known about how the Games will affect local businesses - despite this area having great utility. With this in mind, we pose the following questions:

1. What factors determine whether a local business will see an increase in popularity during the Olympics?
2. Can we use these factors to accurately identify such businesses?

By answering these questions, we hope that local businesses can leverage this information to properly adapt to changes in customer flow that would have been unexpected otherwise.

2 Non-Technical Executive Summary

We find that closeness of a venue to an Olympic event was the most important factor in determining the venue popularity during the Olympics. This is followed by diversity of the venue neighborhood. Lastly, underground station connectivity factors into the venue's popularity. Using these factors alone, we can predict the change in popularity of a venue during the Olympics significantly better than a random guess. Local businesses can leverage these factors to prepare themselves better for benefitting from mega sporting events like the Olympics.

3 Technical Exposition

3.1 Data Collection and Cleaning

In order to measure the popularity of local businesses, we used a dataset of global Foursquare check-ins from April 2012 to September 2013 made public by Dingqi Yang. This dataset contained 33,278 check-ins by 266,909 users on 3,680,126 venues in 415 cities. Given the time

range of this data, we decided to analyze the London Olympics. The venues were not originally mapped to a specific city, but the dataset contained the latitude and longitude coordinates for each venue. Thus, we found the latitude and longitude coordinates for each city in the dataset and mapped a venue to the nearest city based on haversine distance. We filtered out check-in records in London from April 2012 to September 2013 for our analysis, which contains records from before the Olympics to one month after the official end of the Olympics.

The columns in the dataset include: User ID, Venue ID, Timestamp, Latitude, Longitude, Category. Since the information in the Category column is too granular for place diversity analysis (will be discussed later), we used Foursquare API to get the venue category hierarchy[1] and mapped the subtype to the ten major categories: Travel & Transport, Outdoors & Recreation, Shop & Services, Arts & Entertainment, Professional & Other Places, Food, Nightlife Spot, Residence, Event, College & University. In the meantime, We also use the subtype to analyze the attractiveness of venues. For example, within the Food major type, the subtype will be Coffee Shop, Italian Restaurant, Ramen House etc.

3.2 Initial Exploration

So working with checkin.csv, it contains total check-ins 332472, for 54278 venues by 11220 users which constitute from April 2012 - September 2013. As we aim to focus on the London Olympics to confine our scope to London City, we filtered out the check-ins for London City.
The dataset head looks like as shown in Fig

	User ID	Venue ID	UTC time	Timezone offset	Latitude	Longitude	Category	City
0	262915	4aec9f4bf964a52091c921e3	Tue Apr 03 18:00:39 +0000 2012	60	51.498044	-0.090546	Pub	London
1	129494	4aec9f4bf964a52091c921e3	Sun Apr 08 18:18:30 +0000 2012	60	51.498044	-0.090546	Pub	London
2	262915	4aec9f4bf964a52091c921e3	Tue Apr 10 18:16:59 +0000 2012	60	51.498044	-0.090546	Pub	London
3	129494	4aec9f4bf964a52091c921e3	Wed Apr 11 20:56:49 +0000 2012	60	51.498044	-0.090546	Pub	London
4	129494	4aec9f4bf964a52091c921e3	Sat Apr 14 20:29:13 +0000 2012	60	51.498044	-0.090546	Pub	London

Popularity Score Calculation:

So, we assume a venue is popular for a given granular Category, when the number of check-ins done by the user are more for that venue.
In order to accurately estimate the popularity of a venue w.r.t to other venues of the same category, we compute the number of check-ins done by users for say Category ‘Pub’ over the period of time.

[1] https://developer.foursquare.com/docs/build-with-foursquare/categories/

Once, we have the per category_type total check-ins we divide this by # check-ins made for a venue given that type.
P denotes of a venue, C denotes check-in for that venue, popularity score is computed as:

Popularity_score(venue | category = category_type) = C(venue) / sum(C(venue | category = category_type))
This gives us the probability in ranging from [0,1], which acts as an attribute in determining which venues of different types were popular for a given period.

Analysis of Pre, Post and During Olympics:

To successfully retrieve , london_Checkins.csv for any start and end period, we first convert UTC time into a timestamp. Then using a custom made function, we pass the days we want to do

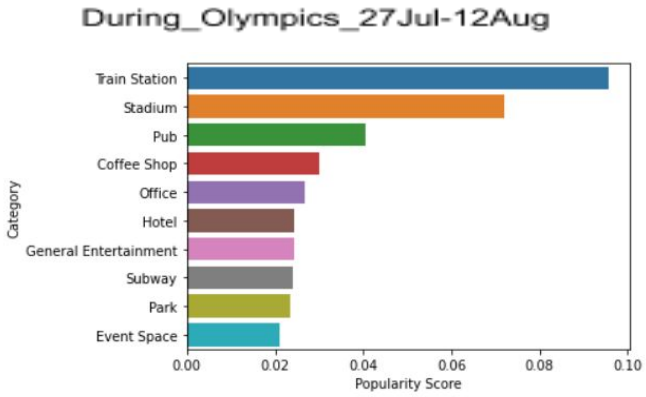
```
def getSpecificTrends(df,type_,days_before=None,days_after=None):

  #during olympics check for UK or London checkin trends
```

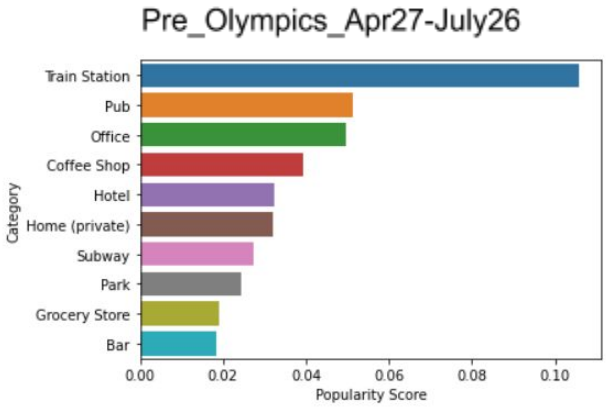
Our analysis. In the Fig 2 shown below, contains analysis of pre olympics and post olympics of 90 days. Apart from that it also contains the trends during the Olympic (27th July - 12th Aug)

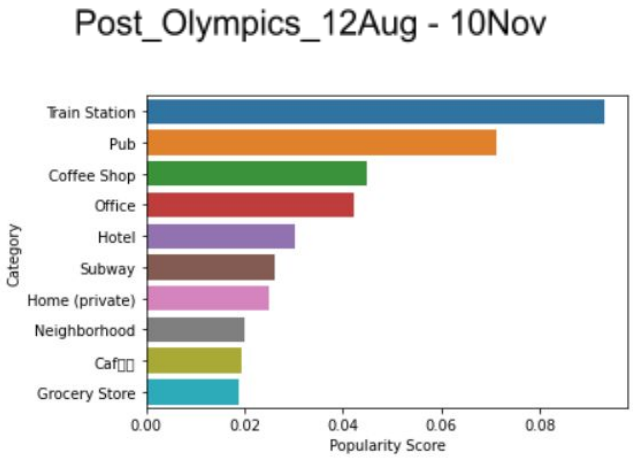
It is clearly evident that, during the Olympics the check-ins shoot up for Stadium. One more interesting thing to notice is, people are more engaging in general entertainment (which include games venues, parks).
Other than that, Train Stations have been on the top for all period, one of the believes could be many office employees, travelers, international visitors would explore the city by using subways or trains. So stations importance is one of the important factors in determining the venue importance as well, which is explored in-depth in the later section.

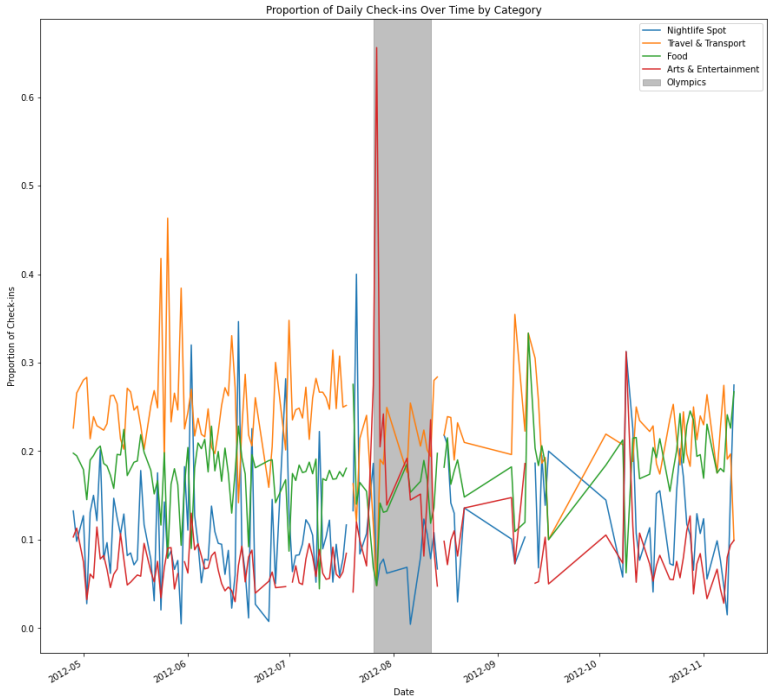
In the inference, while olympic check-ins arose for general activities, parks, stadiums people are coming on the streets of London and that could lead to rise in local businesses.

Exploratory Data Analysis 2:
We plotted time series as a function of check-in made on various venues. Now, we have not used granular categories like (Coffee shop, Sandwich Place, Grocery Store) but instead encapsulated them into the main category called “Food”.
We can notice here for “Arts & Entertainment” which includes all sporting events and activities category were latent for pre-olympics and as soon as Olympics comes in, there has been a sudden exponential increase in check-ins. Post Olympics, again that category seems to slow down. This shows potential areas of research on how the local business can be impacted.









Check-in Trends Over Time

3.3 Problem Formulation

We formulate our problem as a binary classification task aiming to predict whether or not the total number of check ins at a venue increased during the Olympics period as compared to an earlier period of same duration.

Denote the number of check ins at venue v during Olympics (25 July 2012 - 14 August 2012) as $checkin_{after}(v)$ and the number of check-ins at venue v before Olympics (4 July 2012 - 24 July 2012) as $checkin_{before}(v)$.

Then, we aim to predict if $checkin_{after}(v) > checkin_{before}(v)$.

We will use place and venue interchangeably in the below report.

3.4 Feature Engineering

We use two sets of geographic features to assess the spatial information about the venues: proximity features and quality features. Olympic Distance is a proximity feature, while Subway Pagerank, Diversity Entropy, Jensen Quality are quality features. We will use these features to test our hypotheses about the underlying factors contributing to the success of local businesses during the Olympics:

a) Being close to travel stations should contribute to the popularity of a venue, especially during the Olympic event period. (Station Connectivity)
b) The distance advantage of the place being close to Olympic event locations should be a major factor that affects the popularity of the venues during the event. (Olympic Distance)
c) A more diverse area offers many different kinds of activities and users will be attracted to stay in the area for a long period of time and check out more places around the neighborhood (Diversity Entropy).
d) The attractiveness of the venue. Some places tend to attract other places (theme parks attract food trucks) and some tend to repel. If the venue we are interested in tends to attract many other businesses compared to other venues of the same type, we assume that it's more attractive.

We will introduce each feature one by one below.

3.4.1 Station Connectivity

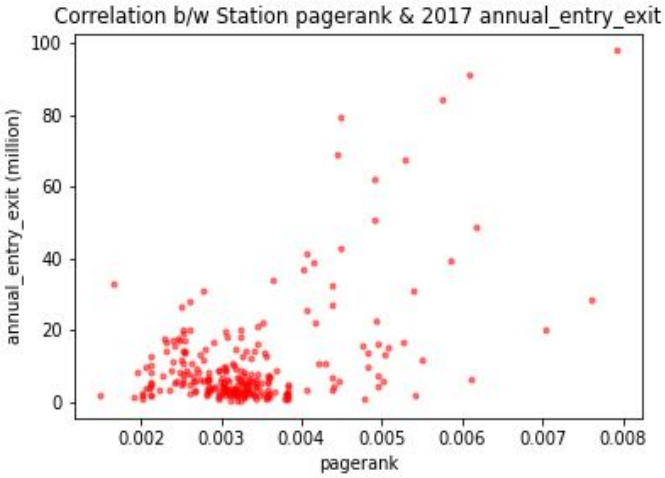
Station Connectivity captures the connectivity of a venue to important subway stations. The London Underground (popularly known as the 'Tube') is a rapid transit system serving Greater London and some adjacent counties. With a daily ridership of appx 5 million[2], it is an extremely cheap and important mode of commute. The popularity of a venue should be affected by its closeness to the nearest station as well as the importance of the station. To measure the importance of a station, we model the station network as an undirected graph and run Pagerank on it[3]. PageRank is an algorithm used by Google Search to rank web pages in their search engine results. However, it can be generalized to apply any graph to measure node importance.

[2] https://en.wikipedia.org/wiki/London_Underground
[3] http://markdunne.github.io/2016/04/10/The-London-Tube-as-a-Graph/

The popularity of a venue *v* should be directly related to the importance of the nearest station *s* but inversely related to the distance from *s:*

$$StationConnectivity(v) = \frac{Pagerank(s)}{Distance(v, s)}$$

It might be interesting to visualize how station pagerank correlates with the annual entry-exit count of the station. We use london_underground_activity.csv to extract annual_entry_and_exit count for each station for the year 2017 and plot its correlation with station pagerank. We report a positive Pearson correlation of *0.50.*



3.4.2 Olympic Distance

Olympic Distance measures the mean haversine distance between a venue *v* and the *k* closest Olympic venues, which were taken from the london_ticket_sales.csv file originally provided. This feature is an extension of the Olympic Distance feature used by Georgiev et al, which was calculated as the haversine distance between *v* and the nearest Olympic venue. We thought that taking the mean distance to the *k* closest Olympic venues would give a more accurate representation of the "connectivity" of *v* to nearby Olympic venues. For the sake of our analysis, *k* was set to 3, but this was a fairly arbitrary choice that can be fine tuned as a next step.

3.4.3 Diversity Entropy

This feature aims to assess the quality of the neighborhood surrounding our target venue *v* in the aspect of diversity (heterogeneity). We calculate this metric in terms of the types of places inside the neighborhood. The granularity of the type used here is the major type.

We classified all venues into ten major types used by Foursquare and we use letter *T* to denote all major types and letter *t* to denote one of the major types.

Given a venue v, the number of venues (including itself) within the neighborhood of *v* is denoted by *N(v, r)*. Similar to *N(v, r)*, the number of venues of type t within the neighborhood of *v* is denoted by $N_t(v, r)$. To measure the neighborhood diversity, we use this entropy metric calculated below which has its root in information theory:

$$-\sum_{t \in T} \frac{N_t(v,r)}{N(v,r)} \cdot \log \frac{N_t(v,r)}{N(v,r)}$$

3.4.4 Jensen Attractiveness

This feature aims to assess the quality of the neighborhood surrounding our target venue *v* in the aspect of target place attractiveness. The category granularity used here is the subtype. Since we use letter *t* to denote one of the major types, we will use subscripts t_v and t_p here to denote the chosen subtype. Jensen (Jensen 2009) introduced a coefficient that quantifies the dependency between two types of places, denoted as k.

$$k_{t_p \to t_v} = \frac{N - N_{t_p}}{N_{t_p} \cdot N_{t_v}} \sum_{q \in P} \frac{N_{t_v}(q,r)}{N(q,r) - N_{t_p}(q,r)}$$

The N is the total number of places in the host city, which we approximate by using the total number of unique venue IDs in London in the check-in dataset. P is the set of places in the host city. $k_{t_p \to t_v}$ is the coefficient between subtype t_v and t_p. Again, similar to the above notation, $N_{t_p}(v,r)$ simply represents the number of venues of type t_p in the neighborhood of venue *v*. The higher the Jensen Coefficient, the more likely that the two types of places tend to attract each other. To illustrate, below is what we got for subtype t_p = University.

Place type (t_v)	$k_{t_p \to t_v}$	Place type (t_v)	$k_{t_p \to t_v}$
British Pub	0.19	Science Museum	24.82
Office	0.23	Airport Terminal	18.88
Home (Private)	0.40	Stadium	9.43
Grocery Store	0.49	Cupcake Shop	8.45
Park	0.75	Farmers Market	6.98

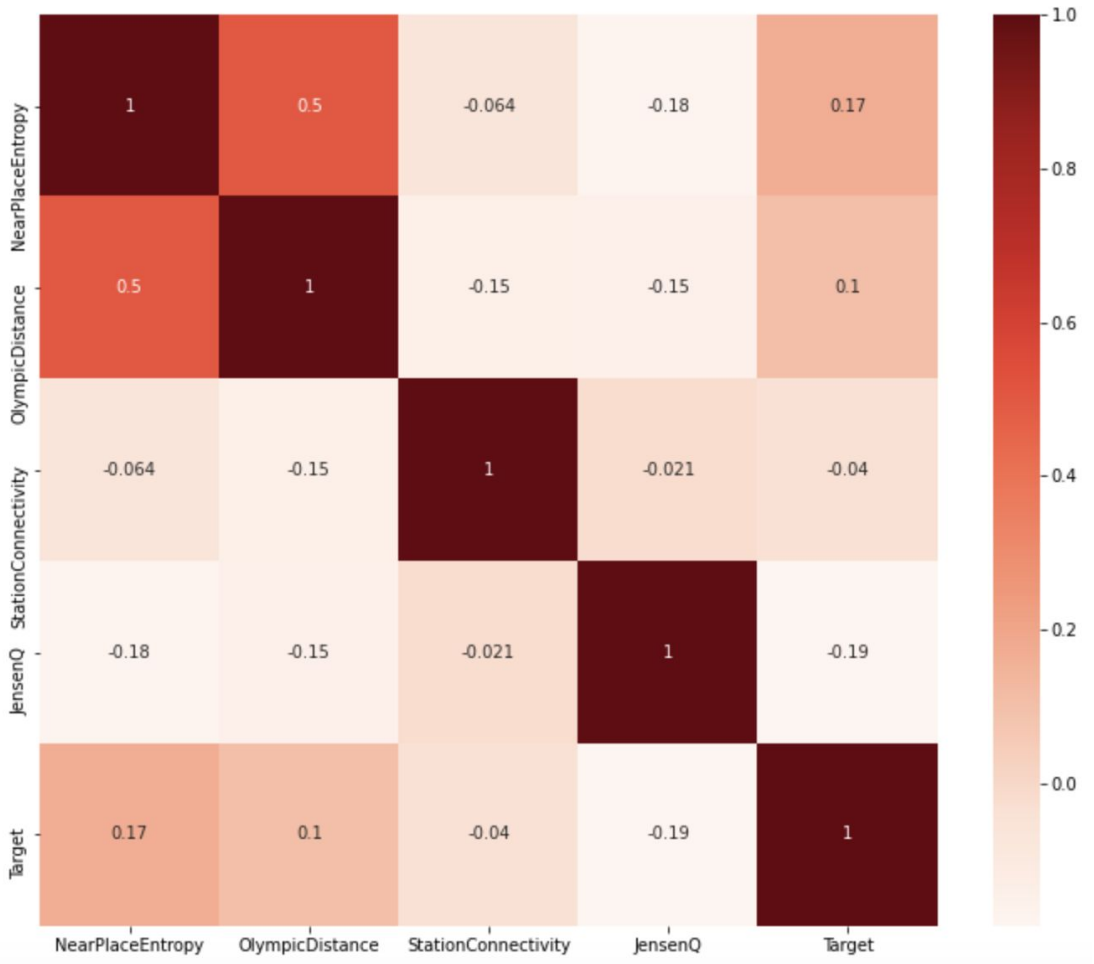
The table shows the lowest (left) and highest (right) Jensen attractiveness coefficient for University.

Again, defined by Jensen, the Jensen Attractiveness is calculated as below. $\overline{N_{t_p}(v,r)}$ denotes how many places of subtype t_p are observed on average around places of subtype t_v. Given a high Jensen attractiveness coefficient k between t_v and t_p, if there are more such t_p places around

venue v compared to other venues, the v neighborhood is considered more attractive. The higher the Jensen Attractiveness score, the more attractive the venue v is.

$$\sum_{t_p \in T} k_{t_p \to t_v} \cdot (N_{t_p}(v, r) - \overline{N_{t_p}(v, r)})$$

3.5 Feature Analysis



With our features, we first performed MinMax scaling to ensure that the features would be suitable inputs for general Machine Learning models. After standardizing the data, we calculated a correlation matrix to identify potential multicollinearity and gain an initial glimpse at the degree to which each feature is correlated to our target variable.

As seen in the heatmap to the right, our features do not appear to exhibit strong correlation with the target variable. One observation is that the NearPlaceEntropy and OlympicDistance have a Pearson correlation coefficient of 0.38. We do not believe this is strong enough to definitely

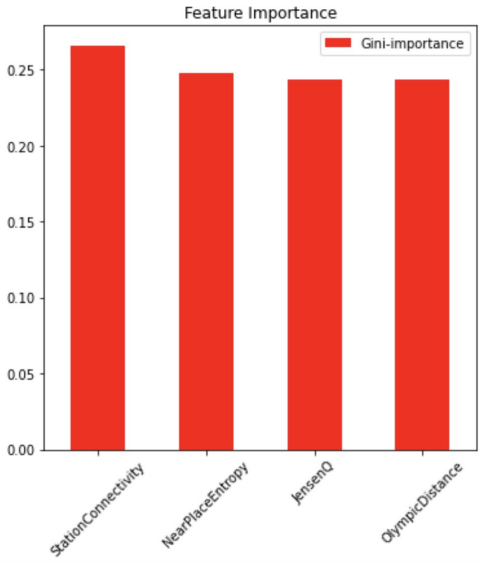
conclude that multicollinearity is present in our data, but it is important to keep in mind when thinking of ways to improve our results.

3.6 Model Results

To analyze the results of our models, we used AUC score as a summary statistic to measure the overall performance of the model.

We created a train/test split of 75/25 and fine tuned the hyperparameters for each model using the training set and 3 cross-validation folds. We then created a soft-voting classifier using the models previously trained. We decided to use soft-voting rather than hard-voting as we believed that incorporating the probability each model assigns to a class will lead to better results than a strict majority of votes.

```
====== Test AUC Scores ======
No Skill:                0.500
----------------------------
SVM:                     0.539
Random Forest:           0.689
Logistic:                0.559
Voting:                  0.664
```

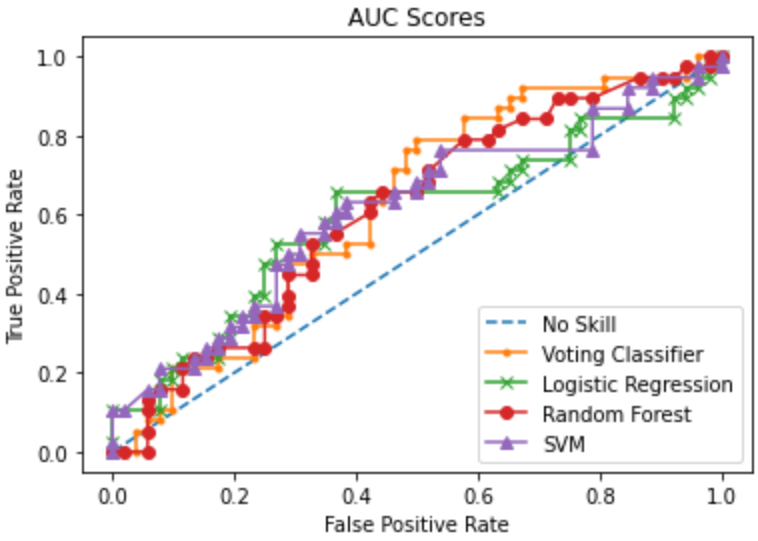


Afterwards, we calculated the AUC scores for each model using the test set. These results can be seen in the table to the left. From this table, we can see that individually, the SVM performed marginally better than the Random Forest on the test set while both performed better than Logistic Regression. Furthermore, we can see that the Voting Classifier performed better than any of the individual models. This result is not surprising as the individual classifiers are rather diverse and should be making uncorrelated errors.

Furthermore, from the Random Forest we are able to see the impurity-based feature importances. It is interesting to note that OlympicDistance has a greater Gini-Importance than NearPlaceEntropy and StationConnectivity. A possible implication from this is that consumers place greater value on “walkability” from the Olympic venue they were just at rather than the

attractiveness of the location or ease of transport. Given this, a future step could be exploring more geographic features similar to OlympicDistance rather than social or mobility focused features like NearPlaceEntropy or StationConnectivity.

Unfortunately, none of the trained models were able to achieve a very impressive AUC score, only being slightly better than random. These results are plotted below. Initially, we believed that the features we analyzed had a fundamental relationship to the popularity of a venue during the Olympics, but upon further analysis it appears the predictive power of the features was less than expected.



3.7 Conclusion

In conclusion, we were able to create a predictive model that could identify businesses that would see an increase in popularity during the Olympics with somewhat reasonable accuracy using geospatial, mobility and user social profile features.

References

Cranshaw, J.; Toch, E.; Hong, J.; Kittur, A.; and Sadeh, N. 2010. Bridging the gap between physical location and online social networks. In Ubicomp.

Dingqi Yang, Daqing Zhang, Bingqing Qu. Participatory Cultural Mapping Based on Collective Behavior Data in Location Based Social Networks. ACM Trans. on Intelligent Systems and Technology (*TIST*), 2015.

Dingqi Yang, Daqing Zhang, Longbiao Chen, Bingqing Qu. NationTelescope: Monitoring and Visualizing Large-Scale Collective Behavior in LBSNs. Journal of Network and Computer Applications (JNCA), 55:170-180, 2015.

Jensen, P. 2006. Network-based predictions of retail store commercial categories and optimal locations. Physical Review E74(3):035101+.

Mehaffy, M.; Porta, S.; Rof’e, Y.; and Salingaros, N. 2010. Urban nuclei and the geometry of streets: The ”emergent neighborhoods” model. Urban Design International 22–46.

P. Georgiev, A. Noulas, and C. Mascolo. Where businesses thrive: Predicting the impact of the olympic games on local retailers through location-based services data. In ICWSM’14.

Zhou X, Hristova D, Noulas A, et al. (2018) Evaluating the impact of the 2012 Olympic Games policy on the regeneration of East London using spatio-temporal big data. Epub ahead of print 5 July 2018.