Urban Data Technical Description

Chunyu Yang (cy356), Ming-Han Tsai (mt627), Yifu Liu (yl896)

We want our Machine Learning model to be an agent to help us to choose an ideal spot to open a fitness center, and also provide us the estimated visitors per month in different spots.

We retrieved the datasets from Safegraph POI (Point of Interest) data to get the information about fitness centers, such as geolocations, raw monthly visitors, brand names, etc. We used the geolocations to get attributes like amenities, restaurants around the geolocations by OSMNX, which is a Python package that lets you download, model, project, visualize, and analyze real-world street networks and any other geospatial geometries.

We have 3 attempts for our model, the first two were not successful since we found the r-square values were negative and there was more space to improve. For the third attempt, we used the Random Forest model to fit and predict the datasets and four spots, which were created by the quantitative methods. The r-square value was 0.26; Even though there is more space to improve the value, we still have found the most important features for the model and verified the result.

Based on the results, we know that the most important feature is the Bakery, which means the count of bakeries around a spot in Manhattan is important to monthly visitors of fitness centers. We verified that the number of visitors increases with the increase in the number of nearby bakeries within 1 km, which proves the correctness of our model. The result is shown below.

Estimated Visitors	164	161	160	148
Bakeries around with 1 km	27	27	22	12

We will choose other models to improve the result and use other physical methods like model validation or field verification to improve the result. However, the current model gives us a reasonable prediction of the monthly visitors for different spots.