Urban Data Final Project Report

Opening a Fitness Center - Site Decision Based on Pedestrian Predictions

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

Cornell Tech, Cornell University

Urban Data, Fall 2021

Emma Pierson

Dec. 7, 2021

Opening a Fitness Center

Site Decision Based on Pedestrian Predictions

Abstract

To select a new site for small businesses, the number of people and the location play a crucial role in the game. In this project, we use visitors' information from SafeGraph, and geospatial data from a Python package, OSMNX. The SafeGraph Database provides detailed information about the number of visitors to certain sites, including where do they come from, how long do they stay in a single block or area, and the rush hour of the area. On the other hand, OSMNX enables downloading geospatial data from OpenStreetMap API, with modeling, visualization, and analysis of real-world street networks and various amenities. With this data, we train a regressor based on selected features leveraging machine learning techniques, and select a few sites within Manhattan, and potentially expand the model to other cities. The result will be compared with the optimal sites that are chosen by a professional real estate broker, to validate the accuracy of the model.

Keywords: Machine Learning, site selection, real estate, linear regression, decision tree, random forest, normalization, decision making

Introduction

We want to open a Fitness Center in Manhattan, and have already short-listed our options to merely four of them. We would like to build a model to know which of the four spots on the list would provide us the best shot for our new businesses, based on data from SafeGraph and OSMNX. This problem occurred frequently in Ming-Han Tsai's past career as a professional real estate consultant. His customers often found it insecure to decide on opening sites and ultimately relied on their experience. Our motivation is to construct a machine learning model using quantified data to underpin the decision-making process and provide robust reasons for our decisions. Once the model is built and validated by professionals, we can further expand it by applying it to various other business types, such as restaurants or grocery stores.

Method

Based on ordinary site selection logic, a place would be more suitable for a fitness center if there are more residential houses or offices in the nearby areas. There might be other neighboring facilities that affect the success of a business and we would like to gather all those information. We extracted the location of all of the fitness centers in NYC from SafeGraph's POI data and used the coordinates in these data to get each location's nearby features as the major determining factors for our model.

SafeGraph is a company providing a powerful dataset that contains numerous categories of point of interest (POI) to enable location-based research. The category point of interest they provide includes industrial, healthcare, retail, parks and leisure, etc. The dataset we exploit contains detailed information of POIs of small businesses, such as restaurants, fitness centers,

and grocery stores, with additional details such as latitude and longitude, operating hours, the popularity of nearby sites.

OSMNX is a powerful python package that, given a particular location, indicated by latitude and longitude or simply address, the API will return a set of amenities that approximate the location. We use this API to enlarge the independent variable that we have, aiming to obtain a more accurate model rather than simply using data from SafeGraph.

We used two regression models, linear regression and random forest regressor. Linear regression is an approach for modeling the relationship between a scalar response and one or more dependent variables. In this case, the independent variables are numerical data from SafeGraph and OSNMX API, and the dependent variables are the popularity of the given location. A random forest classifier examines the key features that significantly affect pedestrian demography. The features we examined are the nearby stores, building types, and other infrastructures that may affect the pedestrian demographics.

We split the dataset into training and testing sets to see which feature combination makes the best prediction, focusing on the data of fitness centers. To preprocess the data, we first drop NA values, and normalize the numerical features, and grid-search the optimal set of parameters including more attributes into our dataset to improve the prediction accuracies. Besides, we also fine tune the hyperparameters such as to ensure that the model does not overfit the training set.

Most importantly, the features we choose initially are based on empirical rules that are intuitively reasonable - for example, a fitness center would achieve higher revenue if located in a residential area rather than that in a commercial district - and then further adjust our model based on the assumption.

Experimental Analysis

First attempt - a few intuitive attributes

We have implemented linear regression on the dataset and the r2_score on the validation set is -0.44. It is not a satisfactory result for now. In addition, we used a random forest classifier to train the dataset, but the score for the training set was 0.97, and the score for the validation set was -0.07. This indicates that the model heavily overfitted the training set, leading to less accuracy on data points that are new to the prediction function.

We examined the coefficients of the attributes and realized that there were many unreasonable relationships between the attributes and visits. For example, the school's coefficient was -2.7, meaning there is a negative correlation to the number of visits, which is counterintuitive. We could not think of an explanation to interpret the findings. This could be caused by the fact that we did not normalize the dataset.

Second Attempt - normalization and much more attributes inclusion

Given the unsatisfactory performance and overfitting at the first attempt, we added more attributes to the dataset to the model, such as public transport, which includes subway, subway entrances, bus stops, and railways, etc. Besides, we also perform normalization on our dataset.

In our second attempt, Linear regression R squared -0.4 was the best score we got and for the random forest, we got 0.15 R square on our validation set.

Third Attempt - every attribute available and grid search for parameter selection

In our third attempt, we added every OSMNX tag into the model and used grid search to select the best performance parameters for the random forest model, such as max_depth, min_samples_leaf, max_leaf_nodes, max_leaf_nodes, etc. At last, we got 0.17 R square for Linear regression and 0.26 R square for random forest on the validation set.

Conclusion

We have predicted the four spots that we derived from the qualitative methods, three of them are around the boundary between mid and lower town of Manhattan, and one of them is on the right side of central park. For each spot that we have chosen, we found all attributes (such as amenities, sports centers, shops, etc) within 1km around the spot. As the predicted result, the random forest model gives us the estimated monthly visitors from 164 to 148, which is higher than the average monthly visitors of 119, and the median of 81. The third quantile of monthly visitors in our original dataset is in the range of estimated monthly visitors from our model.

In general, we believe facilities like subway and bus stations, restaurants will be several important features in our model. However, one counter-intuitive fact from our model is, the model thinks bakery count around a fitness center is important, which has a 0.4 coefficient rate, and the other features that exert a measure of our model are museums, sport information centers, and artwork centers. It's hard to explain why the model gives us unexpected and unexplainable results, but when we were trying to find the bakery counts around the spots that we chose, we found the bakery counts are positively correlated with the estimated monthly visitors, which shows the correctness of our model.

Future improvements

We can try to improve our model and results in several ways. First, we want to acquire more detailed datasets to fit our model. New datasets may include features like open hours, areas, rental price of each fitness center, which not only give us a better understanding of fitness centers but also give the model more features to analyze. Second, another choice of the model may be the option. XGBoost tree is another option for us. The idea of the XGBoot algorithm is to make sure the current tree will not have the same mistakes as the previous three, but learn the advantage from it and its next step improves the performance. The previous results are corrected and the performance is enhanced.

For the model we currently have, we need to do a field investigation to verify whether bakery counts are positively correlated to monthly visitors of fitness centers, to prove the correctness of the model further. Also, we need to do model validations to verify the correctness of the model.

References

Boeing, G. 2017. "OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks." Computers, Environment and Urban Systems 65, 126-139. doi:10.1016/j.compenvurbsys.2017.05.004

Retail site selection checklist: 7 steps for choosing a new location. Places Data & Foot Traffic Insights. (n.d.). Retrieved November 19, 2021, from https://www.safegraph.com/blog/retail-site-selection-checklist.