Explore New Place

JUNSEONG LEE

March 8, 2018

1 Introduction

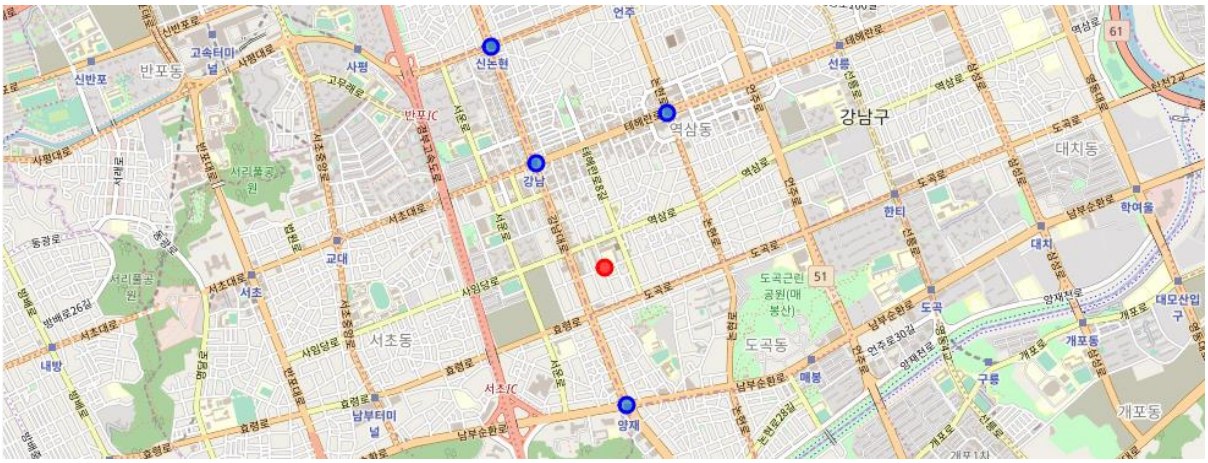
When we are facing situations that we must make a decision in our life, like choosing our university or our first job, the machine learning algorithm can be a good adviser in our life. Some outcomes of choices will be an important decision in our whole life, some will not.

I recognize that my most frequency situation of making a decision is picking my lunch menu every day. I have been worked 3 years in my last job, which is located in Gangnam, Seoul, South Korea, it means I had had lunch in Gangnam nearby the office at least 700 times.

3 years, it's enough to visit almost restaurants nearby the office. But why I'm still in the struggle for selecting lunch menu? This question is my project's start point. I'm going to use a machine learning algorithm to make a recommender which can help me which restaurant is expected to be given a high rating by me.

In Gangnam-gu, there are subway stations in every intersection, so subway station will be a good landmark to find somewhere in Gangnam, for example, restaurants. There are 4 subway stations, Gangnam station, Yangjea station, Sinnonhyeon station, Yeoksam station, near my office, and I find that I never go to Yeoksam station to have my lunch. Even though Yeoksam station is the second nearest place except for Gangnam station from the office!

In this project, I will make an adviser, who predict ratings of restaurants nearby Yeoksam station, Then I can go to a new place to have my lunch with it. It will keep learning my tastes, not even nearby Yeoksam station, it can predict ratings of restaurants in which I heading to!



2 Data

In dataset for the machine learning algorithm, each row is information of the restaurant. I select distance from the office, price level, ratings by Foursquare users, the number of shared tips and photos by Foursquare users as features of my data. I used Foursquare API to collect data

Foursquare API will provide information about places including non-restaurant, so I need to do feature engineering to remove rows which is useless for this project. First, I drop the duplicated rows by place id, 285 rows remained. Second, I remove rows whose 'categories' is non-restaurant and like café or Bar where we will not go to have lunch. After these two steps, 173 restaurants' data retained

The feature 'categories' in the dataset is important information about restaurants, but it refines restaurants into too many types. For example, Janguh Restaurant category only has one place in the dataset, which actually can be in Seafood Restaurant category. For another, someplace is an American Restaurant, some are Sandwich Place, some are Taco Place, all of these only have less than 2 places in the same category. So I create 'New_categories' to turn 36 'categories' into 9 'New_categories'

	name	categories	lat	lng	id	Station	New_categories	distance
3	불이아 (弗二我)	Chinese Restaurant	37.502396	127.036219	52aa9d0111d25f79bea32b9d	Yeoksam	Chinese Restaurant	0.010907
4	Yang Good (양국)	BBQ Joint	37.502146	127.034727	5703972bcd107356d0a6cee9	Yeoksam	BBQ Joint	0.010252
5	고갯마루집	Korean Restaurant	37.499361	127.039308	4d12cb2a12916dcb3d1fd98a	Yeoksam	Korean Restaurant	0.009911
6	BAS BURGER (바스버거)	Burger Joint	37.499723	127.035882	57c65f00498ed5b609eaf4ef	Yeoksam	Burger Joint	0.008299
10	대우식당	Korean Restaurant	37.502750	127.035137	4b8a4a0ef964a520246732e3	Yeoksam	Korean Restaurant	0.010933

About information shared by Foursquare users, I find that many restaurants have more than 50 tips or photos shared. I think the number of shared tips and photos is very important for the decision, but too many tips are useless. Because when I'm searching for the lunch restaurant in office before lunchtime, I see many blog posts or youtube videos on the internet but no more than 30~40 contents. So I refine the number of shared tips and photos to 5 levels. 1st level means there are less than 5 tips that not enough to use for make decision, 5th level means there are more than 30 tips which can fully help to make a decision. Like the table below.

	id	price	rating	photos_count	tips_count	Photos_count_level	tips_count_level
0	52aa9d0111d25f79bea32b9d	1.0	8.0	84	13	5.0	3.0
1	5703972bcd107356d0a6cee9	2.0	8.3	134	20	5.0	3.0
2	4d12cb2a12916dcb3d1fd98a	2.0	8.4	69	16	5.0	3.0
3	57c65f00498ed5b609eaf4ef	1.0	8.0	30	7	4.0	2.0
4	4b8a4a0ef964a520246732e3	2.0	8.2	169	26	5.0	4.0

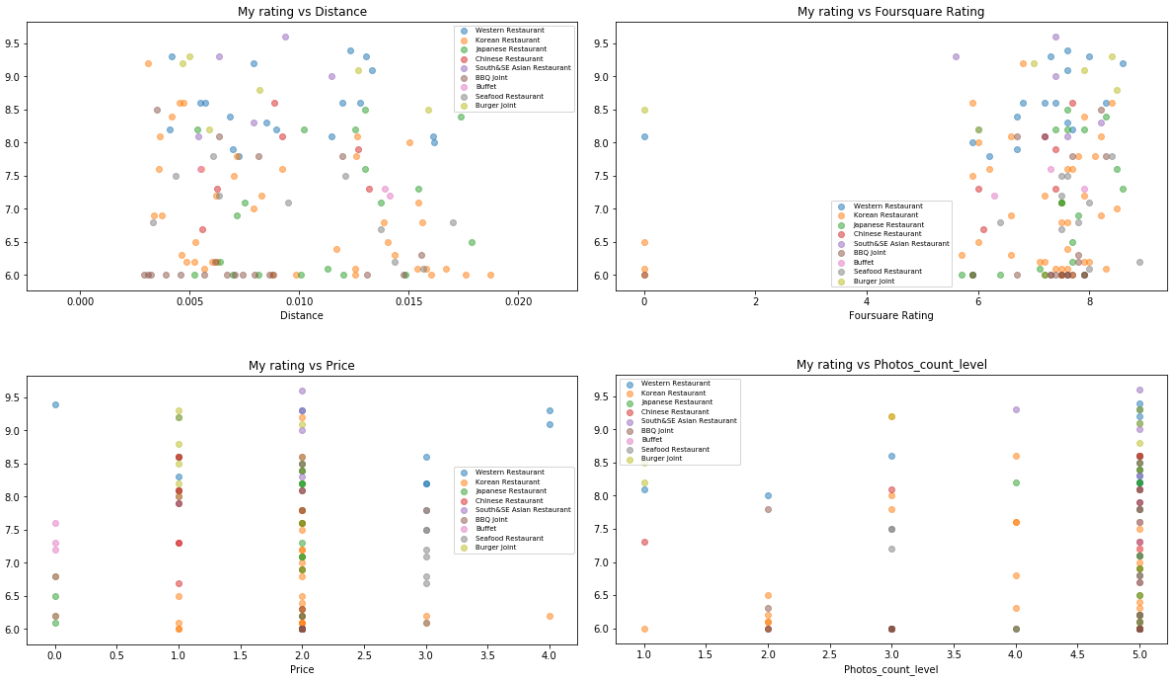
As I mentioned before, this project's purpose is to predict ratings of restaurants near Yeoksam

station, so I need my ratings of restaurants near the office, Gangnam station, Sinnonhyeon station ,and Yangjea station. I do survey myself based on memos, Instagram, blog and more, add My rating to the dataset. What if there are restaurants which I never visited even near the office? I will keep it blank to predict, but fortunately, except for 44 restaurants near Yeoksam station, I had been all of the restaurants in dataset! Thanks to my curiosity about the new taste, 3 years was enough time to visit these 129 restaurants.

Now, the dataset is ready, including information of restaurants ID, distance, categories, price, four square user rating, counts of shared tips and photos, and my rating.

3 Methodolgy

Before doing machine learning, I visualized the dataset to exploratory data analysis. I plot scatters to find out what relationship exists between My rating and each feature. The charts tell me I usually rate high score to Western restaurants and Japanese restaurants. And my ratings of Korean restaurants is lower than ratings of Korean restaurants by Foursquare users, in Western restaurants cases is higher. It means I prefer Western food for lunch to Korean food.

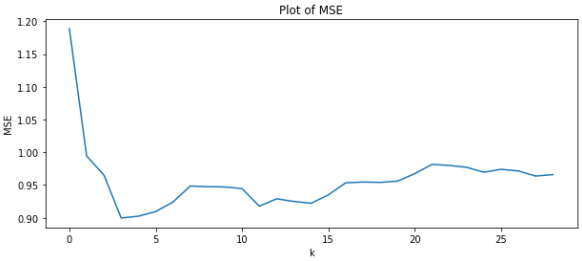


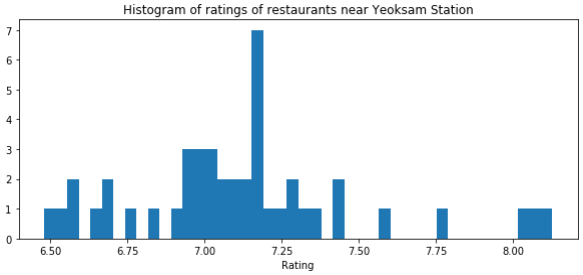
Rating is a continuous feature between 0~10, so I will use KNN Regressor Model instead of KNN Classifier Model. And the feature 'New_categories' is a categorical feature which is not directly suited for KNN model, to solve this problem, I use one hot encoding method to make 'New_categories' usable.

After dropping features like restaurant name, latitude, longitude, nearby station names, and categories only numeric feature retained. Now I split the dataset which restaurants is near Yeoksam

station where my rating is 0. The other dataset which including my rating is divided into train set and test set.

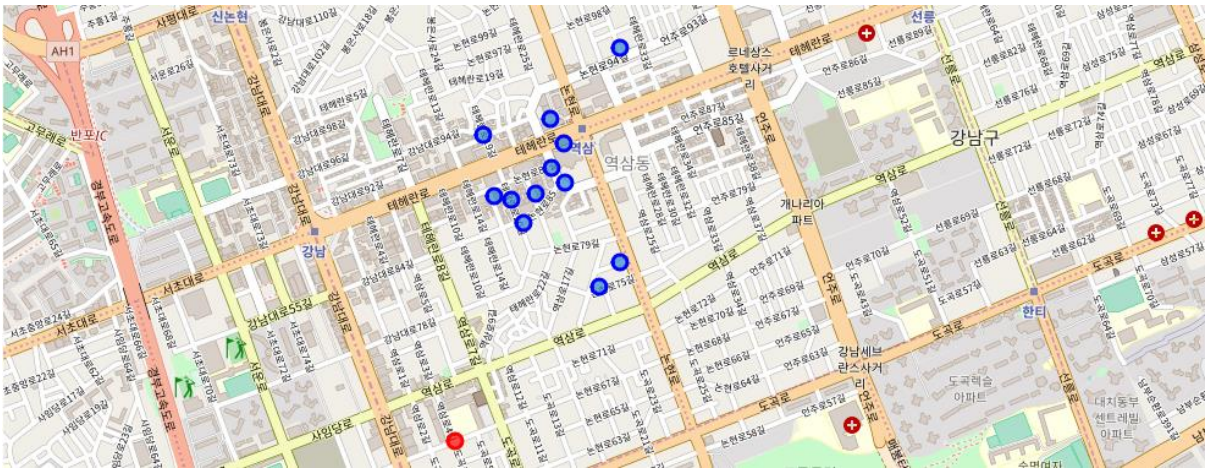
I tried different K and calculate the Mean Squared Error(MSE), plot a line chart of MSE. Then We can see When K is 4, the line chart stop falling start rising slowly, So I choose best k as 4.





4 Result

I predict ratings of restaurants near Yeoksam station, using KNN Regressor Model which k as 4. Drawing a histogram of Ratings can know that most places' rating is between 6.8~7.3. I filte the restaurants which predict rating is higher than 7.25, add markers to map.



The result is similar to what I find during EDA, high score restaurants including Western restaurants, and Burger joints and Japanese Restaurants. So the outcome of this machine learning is, If I go to Yeaksam station to have lunch, I can select the lunch point between makers on the map.

5 Discussion

After making this adviser, I think I can select other places to know where I'm going to eat. But I also worried about someone who loves every food not like me don't prefer to eat Korean food at lunch will be more difficult to recommend restaurants. Then categories is a not important feature for machine learning, maybe the Collaborative Filtering mentioned in the course of Machine Learning with Python in Coursera will be a good solution.

6 Conclusion

This project let me get an adviser which can recommend restaurants for me. It learned my tastes and predict ratings of restaurants. This is my first machine learning project which I choose the topic myself, collected data myself, analyzed myself. It must be many incorrect in the process. If I go to the new place like Fukuoka in Japan or Beijing in China, this first version of the algorithm may not work well, but I believe it will give me more insight of what I to complement. I will keep exploring new places!