Jin Amelia Choi

Applied Data Science Capstone

July 18, 2020

Restaurants of Seoul - ‘Likes’ Prediction

Using Foursquare API and Machine Learning

A street with lots of restaurant in Gangnam district, Seoul

1. Introduction

Seoul, the capital of South Korea, was the 4th largest metropolitan economy of the world with a GDP of US $ 635 billion in 2014 after Tokyo, New York City, and Los Angeles. Seoul has a population of 9.7 million and forms the heart of the Seoul Capital Area with the surrounding Incheon metropolis and Gyeonggi province. So that, there are so many shopping areas which are Dongdaemun Market, Myeongdong, Insadong, and Gangnam district. Of course, it's so activated for commercial areas.

For people in Seoul who want to open their restaurants, knowing ahead of time the potential social media image which they can have would provide an excellent solution to the ever-present business problem of uncertainty. In this case, the uncertainty is regarding the performance of social media presence.

We can mitigate this uncertainty through leveraging data gathered from Four-Square's API, specifically, we can scrape ‘likes’ data of different restaurants directly from the API as well as their location and category of cuisine. The question we will try to address is, how accurately can we predict the amount of ‘likes’ a new restaurant opening in this region can expect to have based on the type of cuisine it will serve.

Leveraging this data will solve the problem as it allows the new business owner (or existing company) to make preemptive business decisions regarding opening the restaurant in terms of whether it is feasible to open one in this region, expect good social media presence and what type of cuisine. This project will analyze and model the

data via machine learning by comparing both linear and logistic regressions to see which method will yield better predictive capabilities after training and testing.

2. Data

2.1. Data Scraping and Cleaning

We are going to use 'Foursquare API' to scrap a bunch of information about restaurants in Seoul. First of all, We define a function to get the geocodes that are latitudes and longitudes using 'Geopy'. And then, we input client id, secret code, and version information to the 'URL' code. Using the 'request' library, we can get the output of the restaurant information.

After making a function that extracts the category of the venue, we categorize the venues and filtering columns that are divided into Name, Categories, Location Latitude, Location Longitude, and ID. Split '.' in the columns and make clean column names. The name of this data frame is 'nearby_venues'.

In this data frame, we extract a list of restaurants to use making 'like list'. The list contains 'likes' which Foursquare users gave. The number of 'likes' show which restaurants of Seoul is popular in the Foursquare users. Code for-loop and extract the list. We finish making 'raw_dataset' to be used for analyzing.

	name	categories	lat	lng	id	likes
0	무교동북어국집	Korean Restaurant	37.567852	126.979753	4ba1a9adf964a5209bc637e3	192
2	Joo Ok (주옥)	Korean Restaurant	37.564705	126.977667	572c2d36498ec2eb8308ade5	41
3	Läderach chocolatier suisse (레더라)	Chocolate Shop	37.568153	126.978265	4caea82aaef16dcbad8ba254	57
6	철철복집	Seafood Restaurant	37.567393	126.981310	4da7cc1393a021ab13bbe742	46
8	Be-Up Coffee (비읍커피)	Café	37.566975	126.979706	5ba2e09d234724002ce3f415	13

2.2. Data Preparation

- **Grouping specific categories**

Korean	Asian	Europe	American	Bar	Casual
Korean Restaurant	Sushi Restaurant	Bistro	Burger Joint	Pub	Chocolate Shop
Seafood Restaurant	Chinese Restaurant	Italian Restaurant	Steakhouse	Hotel Bar	Café
Gukbap Restaurant	Japanese Restaurant		BBQ Joint	Beer Garden	Buffet
Bossam/Jokbal Restaurant	Vietnamese Restaurant		Mexican Restaurant		Dessert Shop
	Indian Restaurant				Coffee Shop
					Tea Room
					Bagel Shop
					Food Court
					Noodle House
					Bakery

- **Adding *'categories_classified'* column**

- **Finding where to bin the '*likes*' data**

- **Creating a function '*ranking*' to bin and rank**

	name	categories	lat	lng	id	likes	categories_classified	ranking
0	무교동북어국집	Korean Restaurant	37.567852	126.979753	4ba1a9adf964a5209bc637e3	192	Korean	1
2	Joo Ok (주옥)	Korean Restaurant	37.564705	126.977667	572c2d36498ec2eb8308ade5	41	Korean	2
3	Läderach chocolatier suisse (레더라)	Chocolate Shop	37.568153	126.978265	4caea82aaef16dcbad8ba254	57	Casual	1
6	철철복집	Seafood Restaurant	37.567393	126.981310	4da7cc1393a021ab13bbe742	46	Korean	2
8	Be-Up Coffee (비읍커피)	Café	37.566975	126.979706	5ba2e09d234724002ce3f415	13	Casual	3
10	Sushi Cho (스시 조)	Sushi Restaurant	37.564491	126.979743	4f82efddd4f208b5bd478f0e	114	Asian	1

2.3. Exploratory Data Analysis

- **The Greatest & Worst Restaurant in Seoul**

	name	categories	lat	lng	id	likes	categories_classified	ranking
65	만족오향족발	Bossam/Jokbal Restaurant	37.563408	126.975884	4ba33ce1f964a520363138e3	222	Korean	1

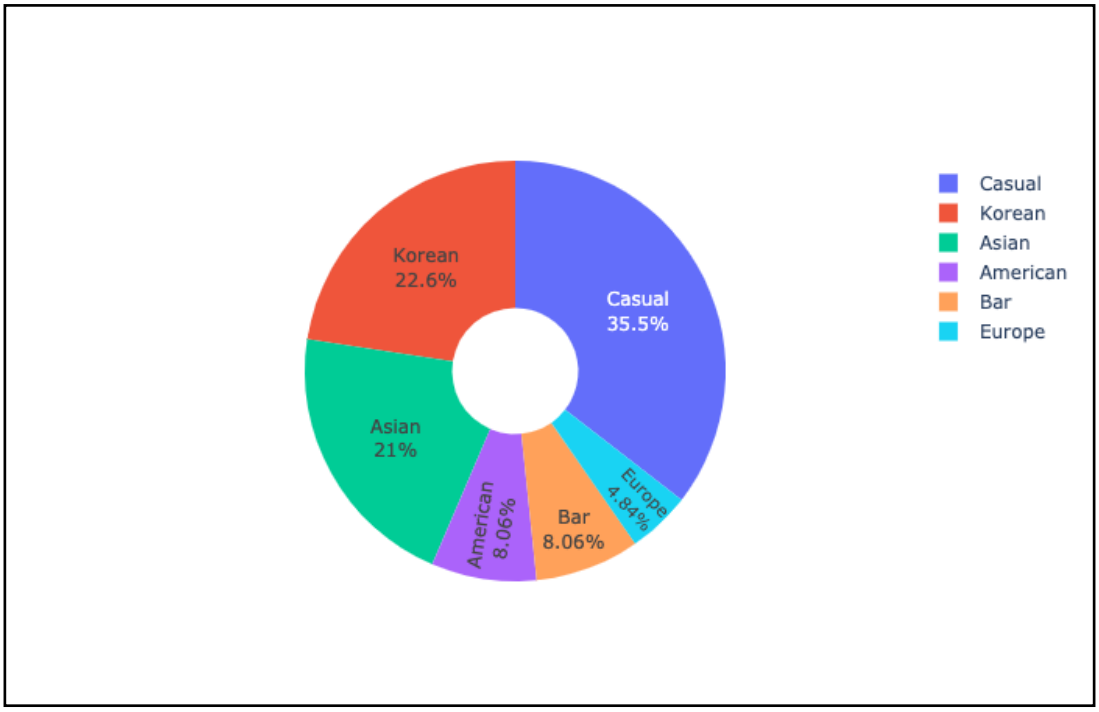
만족오향족발(called 'Manjok Ohyang Jokbal'), which is categorized in the 'Bossam/Jokbal Restaurant', is the greatest restaurant for Foursquare users. - For the people who don't know Korean food - Bossam/Jokbal is made of the pork meat. Bossam is a slice of boiled pork meat, eating with Kimchi and some sauces. Jokbal means in Korean the feet of the pig. Also, it is boiled food for a long time with soy sauce (Ganjang), vegetables, and lots of species. These foods have both chewy and mild taste.

	name	categories	lat	lng	id	likes	categories_classified	ranking
66	비어할레 / Bier Halle (비어할레)	Beer Garden	37.567572	126.981408	4bee9c46e8c3c92840db9892	7	Bar	3

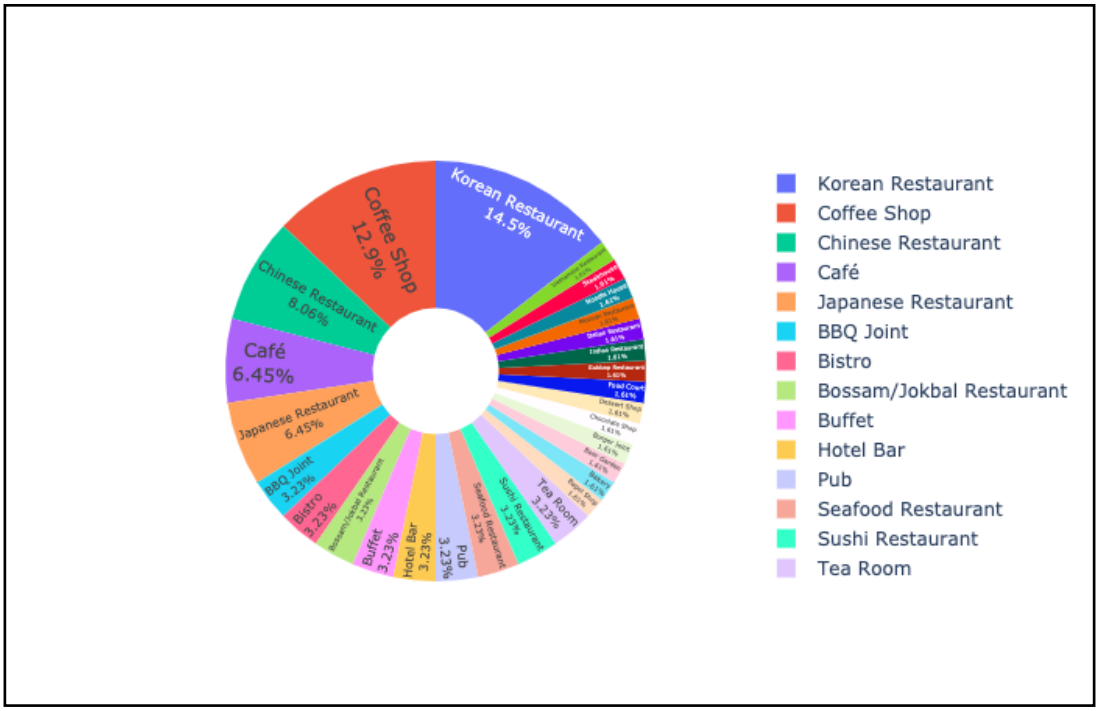
비어할레 (Bier Halle - it has both meanings "Would you take some beer time?" in Korean and 'Bier' representing Germany original beer) is the worst restaurant for users. Maybe some Germans thought that this restaurant's beer isn't original Germany beer. Some Korean wrote reviews very well (for example, it's good for meeting friends and making some great beer time with tasty foods).

- **The Ratio on features in the Dataset**

In this data set, we will discuss both features' ratios, 'categories_classified', and 'categories'. And deeply, we're looking for names of coffee shop/cafe in Seoul. We use 'plotly' library which is rapidly delivering real results and gives great visualization methods in Python, R, and Julia. More information about 'Plotly': check this link.



This graph represents the ratio of 'categories_classified'. Among the restaurants in Seoul, Casual restaurant has the biggest proportion (35.5%) in their categories' classification. And European foods have the least (4.84%) for this classification. The ratio of Casual restaurants is 7 times as big as the European one.



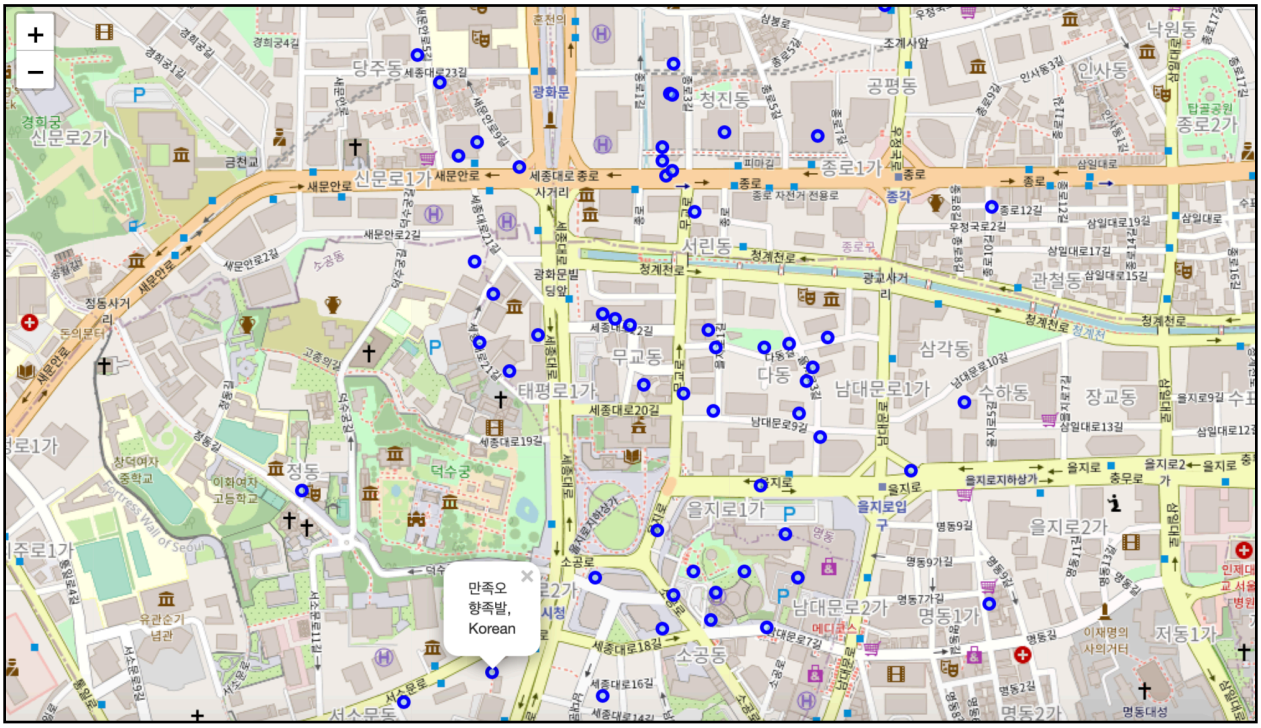
Specifically, among the 'categories' feature, the Korean restaurant category has the largest proportion (14.5%). But there is just a little-term with the second one, the coffee shop category. There have been placed so many coffee shops in Seoul since the 2000s. The increase in the number of coffee shops in Seoul is mainly due to the high demand for coffee. The institute estimated that each adult in South Korea consumes 353 cups of coffee a year. That rate is 2.7 times bigger than the global average of 132 cups a year (Source - South Korea's Coffee Craze, Tae-jun Kang).

	name	categories	lat	lng	id	likes	categories_classified	ranking
8	Be-Up Coffee (비읍커피)	Café	37.566975	126.979706	5ba2e09d234724002ce3f415	13	Casual	3
27	Café MAMAS (카페마마스)	Café	37.567216	126.979188	4b84c56bf964a520a34331e3	79	Casual	1
33	Paul Bassett (폴바셋)	Coffee Shop	37.568019	126.976677	5146a5ece4b088751f847bee	139	Casual	1
46	Paul Bassett (폴바셋)	Coffee Shop	37.570404	126.978832	55efabb5498ef918a753f0f2	46	Casual	2
49	Starbucks Reserve (스타벅스 리저브)	Coffee Shop	37.564463	126.979025	4b5bb69af964a520a31129e3	137	Casual	1
59	CONFECTIONS BY FOUR SEASONS	Coffee Shop	37.570462	126.975299	56175cb7498e144b13689c4b	20	Casual	2
61	Starbucks (스타벅스)	Coffee Shop	37.564000	126.978832	54e4032b498ea9b63904a623	15	Casual	3
64	수수커피	Café	37.571723	126.979010	59afb2389de23b1bcc2de6d8	16	Casual	3
83	Starbucks (스타벅스)	Coffee Shop	37.570267	126.978990	55e24a57498e6b62f70697e0	58	Casual	1
90	Starbucks Reserve (스타벅스 리저브)	Coffee Shop	37.563005	126.974359	5456b3f7498eccfb5b024079	47	Casual	2
92	Executive Lounge Courtyard Marriott	Café	37.561015	126.976850	57738d8b498e5f4d8da00e7c	10	Casual	3
97	Starbucks Reserve (스타벅스 리저브)	Coffee Shop	37.566160	126.983108	590f055c8c35dc6736fce4f5	8	Casual	3

As we were looking for the names of coffee shop / cafe, the result is that Starbucks is the most famous and popular coffee shop in Seoul. What a Starbucks!

2.4. Exploring Data in the Map of Seoul

For the last one, we make a map of Seoul with pointing dots using restaurants' geographic data. To make the map in Python, we use the 'folium' library. Also, using the 'CircleMarker' method in 'folium', we can make labels that represent the name and category information of this point on the map. We set You can choose different shapes and colors of points. Check this link: https://python-visualization.github.io/folium/modules.html



3. Methodology

This project will utilize both linear and logistic regression machine learning methods to train and test the data. Namely, linear regression will be used in an attempt to predict the number of "likes" a new restaurant in this region will have. We will utilize the Sci-Kit Learn Package to run the model.

We can also utilize logistic regression as a classification method rather than a direct prediction of the number of likes. Since the number of "likes" can be binned into different categories based on different percentile bins, it is also potentially possible to see which range of "likes" a new restaurant in this region will have.

Since the "likes" are binned into multiple (more than 2) categories, the type of logistic regression will be multinomial. Additionally, although the ranges are indeed discrete categories, they are also ordinal. Therefore the logistic regression will need to be specified as being both multinomial and ordinal. This can be done through the Sci-Kit Learn Package as well.

4. Result

4.1. Linear Regression

For using the linear regression model, we have to change categorical features to dummy features (called One-hot encoding). In this model, we have two features; 'categories_classified' and 'likes'. 'categories_classified' is the categorical one. So we use

the 'get_dummies' method in Pandas. And the equation result of multiple linear regression is that:

$$y(Likes) = -24.52 * American - 1.32 * Asian + 17.34 * Bar + 17.56 * Casual - 18.65 * Europe + 9.58 * Korean + 54.16$$

In this equation, we aren't concerned about the feature 'rank' - this feature is dealt with Logistic Regression because it's not a regression problem but classification one. To verify prediction capabilities, we checked variance score, mean absolute error(MSE), residual sum of square (RMSE), and R2-score. But these scores were unreal and a little bit awkward (some scores maintain minus number). In conclusion, we decide that the 'categories_classified' feature cannot represent the y score (likes scores) in this linear model.

4.2. Logistic Regression

A multinomial ordinal logistic regression model was trained on a random subsample of 80% of the sample and then tested on the other 20%. To see if this is a reasonable model, its Jaccard similarity score and log-loss were calculated (40% and 1.05 respectively). Although this is not a perfect prediction, a similarity of 40% between the training set and test set is a reasonable result.

Given the modestly accurate ability of this model, we can also run the model on the full dataset. The coefficients show that opening a restaurant in Seoul, opening American or European restaurants, are associated negatively with ‘likes’.

5. Conclusion

Overall, there is a lack of other different features in the 'like' model - both multiple linear and logistic regression models. Even, the multiple linear regression model cannot be established well. We know it for looking at R2-score and MSE/RMSE. The logistic regression model represents 40% accuracy - this score shows how much the features represent this logistic model. If more feasible features are added in the model, the accuracy of the model will be higher than before. We think, there are some features - more detailed information about restaurants and locations.

For the future,

Jin Choi.