Coursera Final Capstone Project

A preliminary analysis of Foursquare ratings of sushi restaurants at New York City

Hugo Tadashi

March 2020

1 Introduction

Sushi is a famous Japanese dish of prepared vinegared rice accompanying a variety of ingredients, such as seafood, vegetables, and sometimes, fruits. Styles of sushi and its presentation vary widely, but the one key ingredient is shari, also known as "sushi rice".

Since its arrival in the United States in the late 1960s, sushi dishes got more and more popular: in 2020 it is estimated that the sushi restaurants industry has a market size (measured by revenue) of $22.1 billion.[1] In fact, in New York City, the most populous city of the United States, there are more than two hundred sushi restaurants with a great spectrum of prices.

Figure 1: Sushi plate. Photo by Ishikawa Ken.[2]

[1] According to https://web.archive.org/web/20200410142904/https://www.ibisworld.com/industry-statistics/market-size/sushi-restaurants-united-states/, accessed on 3rd April of 2020.

https://www.flickr.com/photos/chidorian/

In this project, it will be analyzed if (and how) some features obtained by the Foursquare site can be used to predict the Foursquare rating of sushi restaurants in New York City. As a possible future application, this work could help sushi restaurants owners (or prospective owners) in business decisions (such as credit card adoption) to help their restaurant gain a high score rating in Foursquare and, consequently, increase the publicity of their business.

2 Data acquisition and description

All the data were extracted from the Foursquare site and retrieved by using Foursquare Places API.[3] The next table shows the features that was used as a feature for classifying the rating of a sushi restaurant:

Table 1: Description of data extracted from Foursquare database

Data	**Description**	**Type**
Latitude	Latitude of restaurant.	Quantitative
Longitude	Longitude of restaurant.	Quantitative
Price tier	An integer from 1 (least pricey) to 4 (most pricey).	Categorical
Reservations	1 if restaurant has reservations, 0 otherwise	Categorical
Credit cards	1 if restaurant accepts credit cards, 0 otherwise	Categorical
Outdoor seats	1 if restaurant has outdoor seats, 0 otherwise	Categorical
Delivery	1 if restaurant has delivery service, 0 otherwise	Categorical
Created at	Seconds since epoch when the restaurant was added in Foursquare.	Quantitative
Likes	The count of users who have liked this restaurant.	Quantitative
Number of photos	The number of photos of this restaurant.	Quantitative
Rating	Numerical rating of the restaurant (0 through 10).	Quantitative

An example showing the data extracted from Foursquare for five restaurants is shown in Table 2.

[3] `https://developer.foursquare.com/places`

Table 2: Example dataset

Name	id	Latitude	Longitude	Creation epoch
Momoya	49d991d9f964a5204a5e1fe3	40.7426871	-73.99661748	1238995417
Sugarfish	581a10901df6b32e66ec3a07	40.73895054	-73.98895476	1478103184
Ennju	3fd66200f964a52083e31ee3	40.73736918	-73.99114041	1071014400
Sushi Yasaka	4ea75442f7903beac0782454	40.77942482	-73.98353115	1319588930
Tao	440754f8f964a5204d301fe3	40.76265929	-73.97144108	1141331192

Name	Price tier	Rating	Number of likes	Number of photos	Reservations	Credit cards	Outdoor seating	Delivery option
Momoya	4	9.2	771	469	0	1	0	0
Sugarfish	2	9	453	237	0	1	0	1
Ennju	1	8.1	227	167	0	1	0	1
Sushi Yasaka	2	9	652	494	0	1	0	0
Tao	4	8.8	1214	1790	0	1	0	1

3 Methodology

The dataset was obtained by using the Python request package to communicate with Foursquare API. Since Foursquare API only return at most 50 results per query in a determined latitude and longitude pair, the dataset was obtained by scraping the Foursquare database querying its search venues API for various latitude and longitude coordinates inside New York City. By querying the API for 10 different pairs of latitude and longitude distributed inside the New York City and removing the duplicate restaurants was possible to retrieve 282 different sushi restaurants.

Unfortunately, the retrieved data was not entirely complete:

- 100 restaurants did not have rating and/or price tier information, so these examples were dropped from the dataset.
- 125 restaurants did not have information about credit card acceptance, so it was assumed that these restaurants do not accept payment by credit card.
- 143 restaurants did not have information if accept reservations, so it was assumed that these restaurants do not accept reservations.
- 131 restaurants did not have information if outdoor seats were available, so it was assumed that these restaurants did not have outdoor seats available.
- 128 restaurants did not have information if a delivery service was available, so it was assumed that a delivery service was not available for these restaurants.

Before choosing a model and an algorithm to fit the data, an exploratory data analysis was made to gain a better insight into the relations between the multiple variables of the model and to select the best features for predicting the restaurant rating. The following analyzes were made:

1. A table of the (Pearson) correlation between all the proposed features.
2. A map of New York City with restaurants labelled by its rating.
3. Box plot of the rating versus the price tier.
4. Scatter plot of the rating versus the number of likes and number of photos variables.
5. Box plot of the rating versus other categorical variables.

Finally, ridge regression was used to fit the data to a linear model. The results and the discussion about these results are presented in the next section.

4 Results and discussion

4.1 Exploratory data analysis and feature selection

4.1.1 Correlation between variables

Figure 2 shows the correlation between all variables.

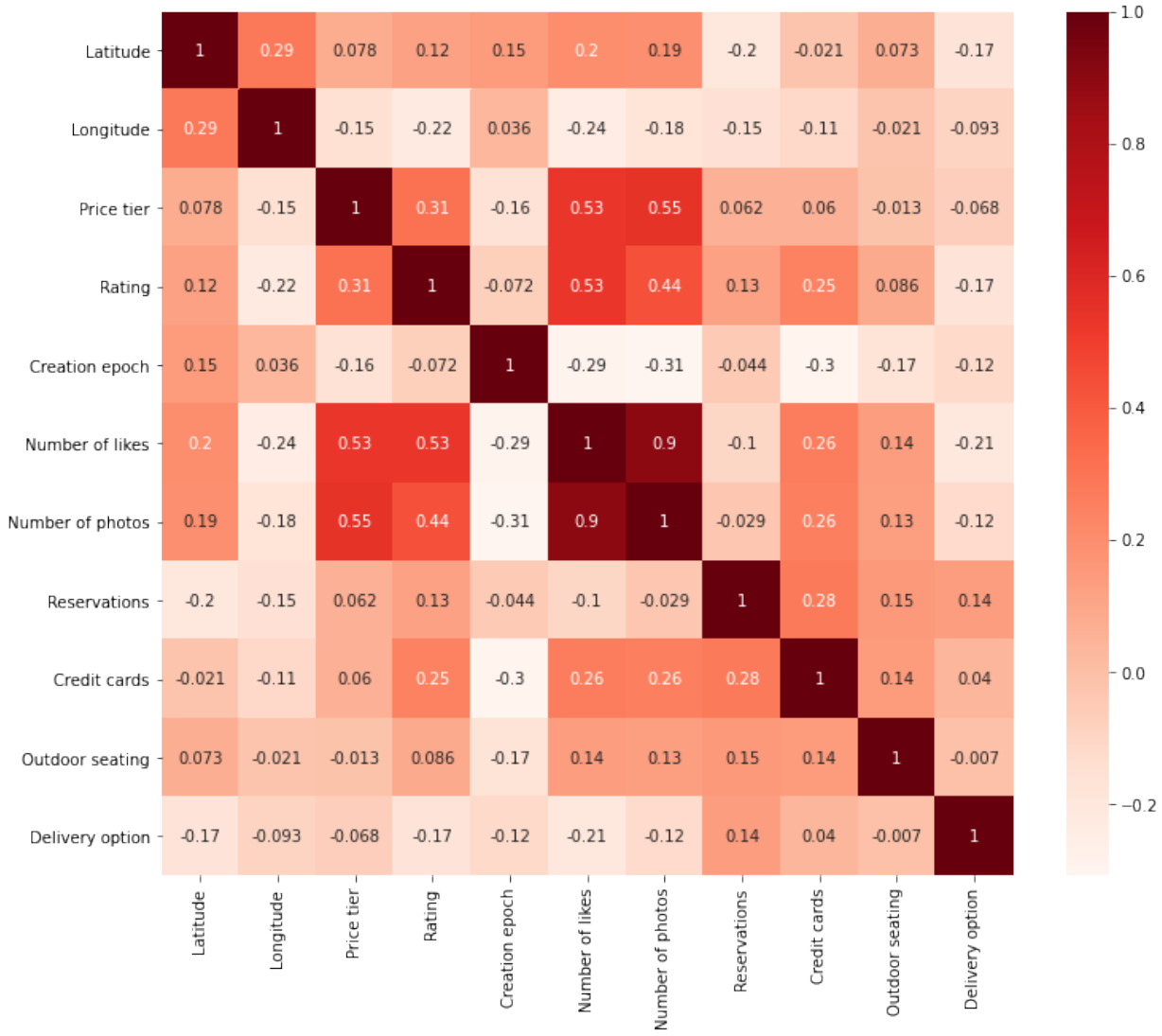

Figure 2: Correlation between all variables.

Figure 2 indicates that the most promising features for predicting the rating are the number of likes and the number of photos in Foursquare. It also shows that the creation epoch, the delivery option and the longitude are unpropitious as features. Nevertheless, with the exception of the creation epoch, the author of this work found interesting to analyze these features even so. The next subsections will analyze these features with more details, with the exception of the creation epoch.

4.1.2 Map plot

The latitude and longitude variables were used to build a map of New York City marked with the restaurants labelled by its rating. A part of the map is shown in Figure 3

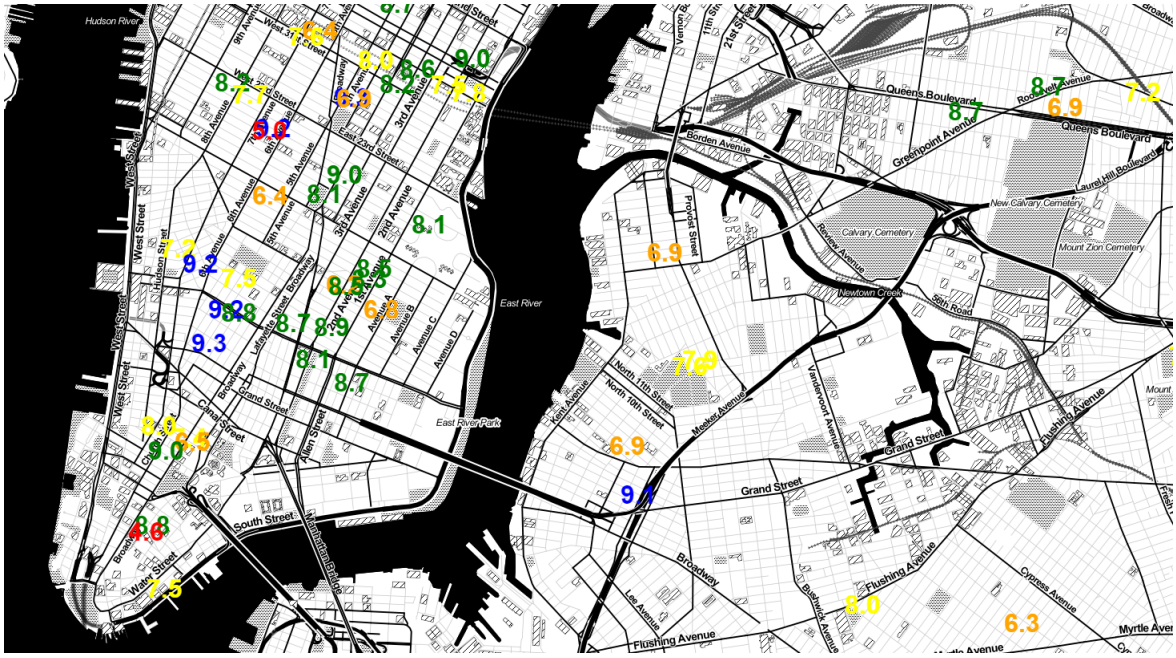

Figure 3: A part of New York City map with sushi restaurants locations and ratings

As can be seen in Figure 3, there is no obvious relation between the rating of the restaurant and its location. This was surprising because it was expected that the location of the restaurant would influence the price tier and consequently, influence the rating of the restaurant. This assumption however is not supported by the retrieved dataset.

4.1.3 Box and scatter plots

The next figures show the box plots between the price tier versus the quantitative variables rating, number of likes and number of photos.

Figure 4 shows the box plot of the restaurant price tier versus its user rating on Foursquare.

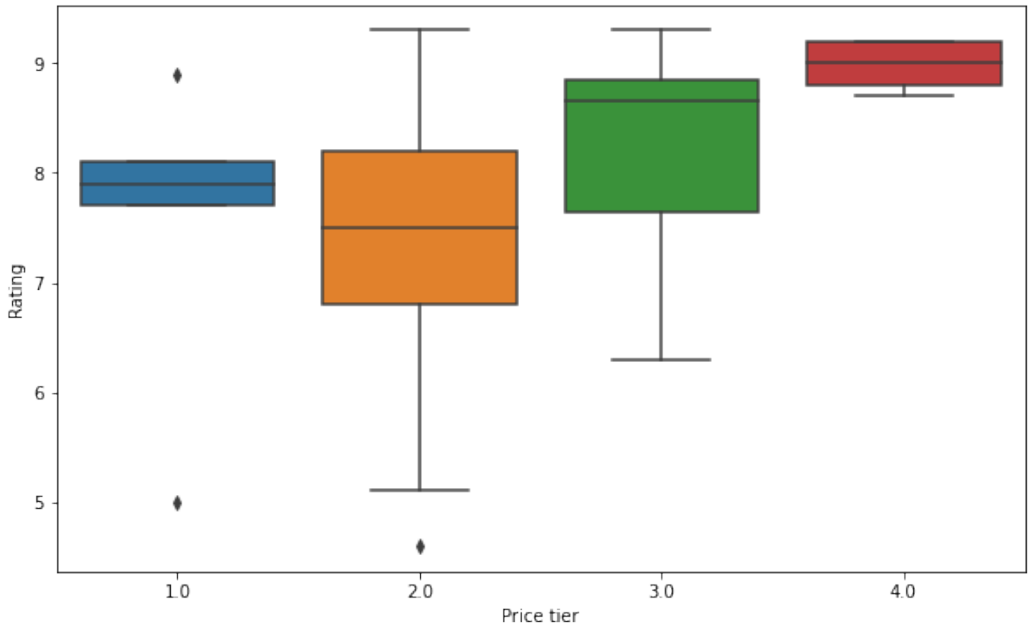

Figure 4: Box plot of price tier versus rating on Foursquare

As expected, the restaurants with the highest price tier (4.0) have a consistent high rating, between 8.7 and 9.2. This is not surprising because restaurants with higher prices, in general, have higher goodwill. Consequently, it should be natural that these restaurants have a good service and a good relationship with customers.

However, it can be seen that there is a great variation of ratings for restaurants with price tier 2.0. Doing a checkup on the dataset it was found that 84.07% of the restaurants have price tier 2.0, while 2.75% have price tier 1.0, 9.89% have price tier 3.0 and 3.30% have price tier 1.0. Thus, while it makes sense to use the price tier as a feature for predicting the price tier, the data obtained was too much unbalanced to be useful.

Figure 5 shows scatter plots of the restaurant rating versus the number of likes and the number of photos in the Foursquare platform. It was decided to make a scatter plot instead of a box plot in order to explore the entire data instead of the summary provided by the box plot.

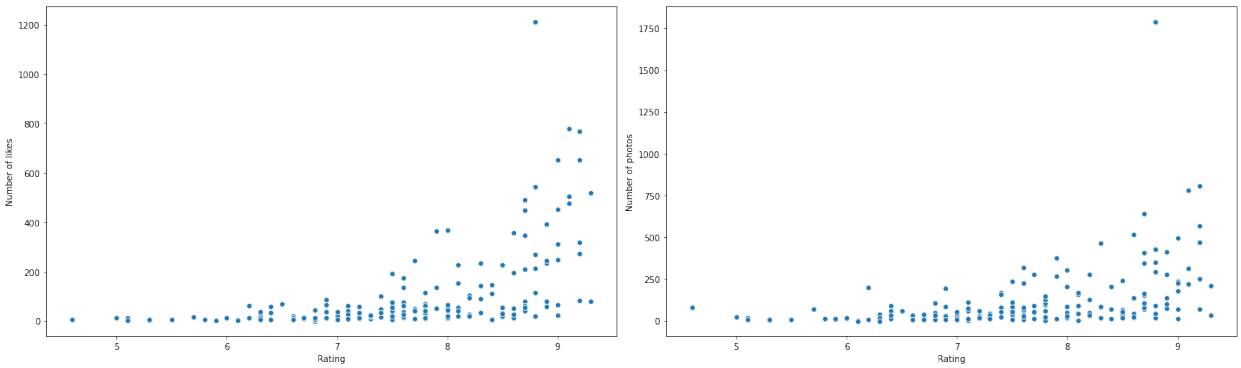

Figure 5: Scatter plots of restaurant rating versus number of likes and photos.

It is natural to suppose that a restaurant with a great number of likes or photos should have a good rating and the scatter plots seem to indicate that these two characteristics are good features to develop a

regression model. Nevertheless, restaurants with a few number of likes and/or photos have a broad spectrum of ratings: this could mean that those restaurants did not have much user engagement on Foursquare and consequently, should be interpreted as a noise for the model. Thus, the restaurants with less than 100 likes or 100 photos were dropped from the dataset, decreasing the total of restaurants to only 39 restaurants. Figure 6 shows the scatter plots after the removal of these restaurants.

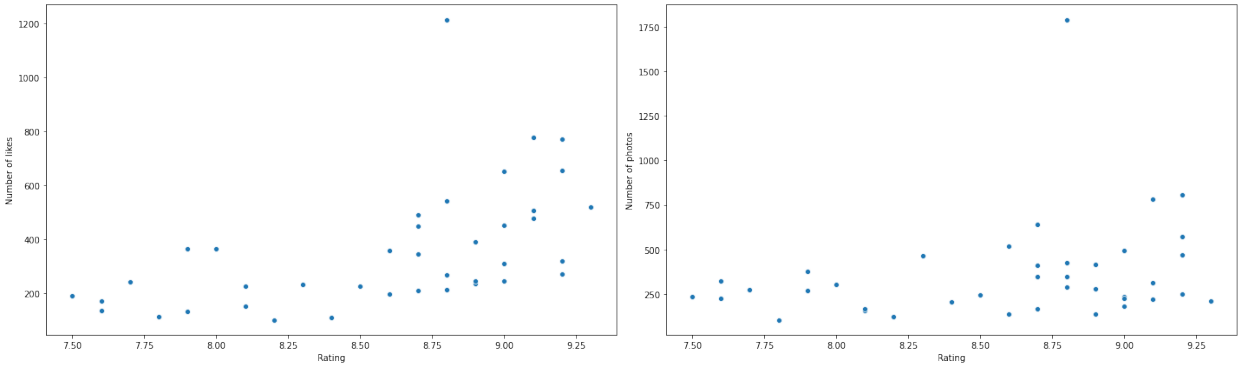

Figure 6: Scatter plots of restaurant rating versus number of likes and photos.

It is also interesting to note the outlier restaurant which has more than 1500 photos, indicating an unusual popular restaurant. This may indicate a restaurant with a distinguishing gastronomic and/or an elaborated architecture which engage the curiosity of a great number of customers.

Figure 7 shows box plots between the price tier and the other categorical variables, that is, if the restaurant accepts credit card, if it is possible to make reservations, if there exists outdoor seating and if there is a delivery service.

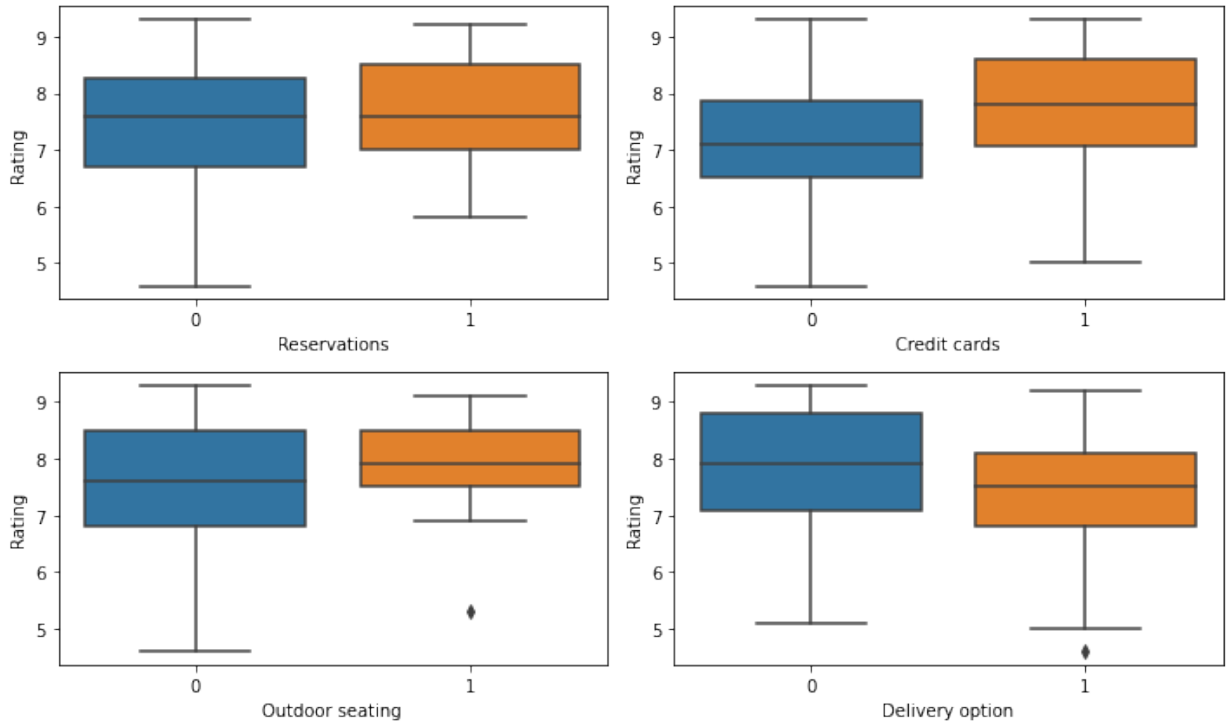

Figure 7: Box plots of the rating versus other categorical variables.

As can be seen in Figure 7, restaurants with lower ratings does not accept reservations nor have outdoor seating. A possible explanation for this is that these services can satisfy some costumers and increase their perception of the rating of the restaurant. Apart from this remark, there is no obvious relation between the availability of any of these services and the restaurant rating in the extracted data, thus these features were not used in the model.

4.2 Proposed regression model

The proposed model for prediction of the rating based on the features of the number of likes and the number of photos is a linear model fitted by Ridge regression, which is a variant of ordinary least squares where the cost function has an additional regularization term with weight α.

Since the final number of samples was small (39 restaurants), the Scikit learn default split of 75% for the training set and 25% for the testing set was not used. Instead, it was decided to split the data such that 90% of these restaurants were used to train a model and the remaining 10% to test and evaluate this model.

The best hyperparameter α was found by brute force searching on a logarithmic space, and for this hyperparameter, the R^2 score was 0.78, indicating a moderate predictive accuracy. Figure 8 shows the linear model obtained by ridge regression with a scatter plot of the dataset.

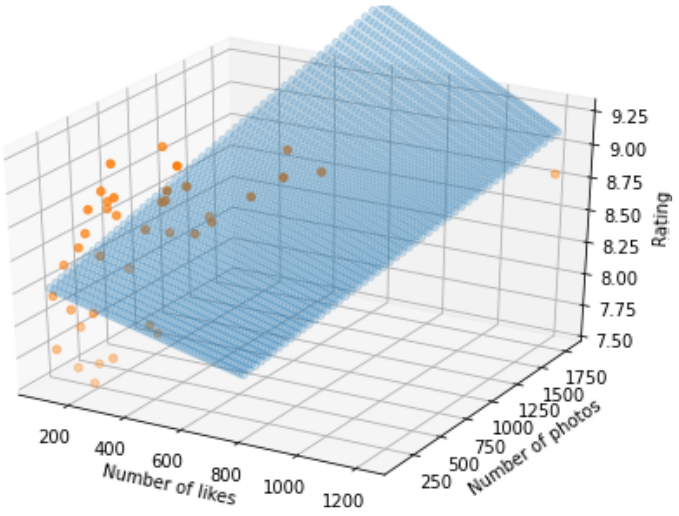

Figure 8: Linear model obtained by ridge regression.

Caution should be made, however, as 90% of the restaurants in the dataset were used to fit the model and only 10% was used to test the model. As will be discussed in the Conclusion, future works with a larger number of samples should be done to validate the obtained model.

5 Conclusion and future works

This project proposed the analysis of the Foursquare ratings of sushi restaurants located at New York City based on the other features available on Foursquare database. Future directions for this work include adding data from other data providers (such as Google Maps) in order to do a more robust analysis and a more trustful prediction model. Another interesting future direction is to add more details to the credit cards accepted (e.g.: only Visa, only Visa and Mastercard, etc.) to see if the brand of credit card influences on the rating. Moreover, sentiment analysis algorithms could be used to automatically classify the reviews in Foursquare as positive and negative reviews, giving an interesting feature to add to the model. Finally, recent works on deep neural networks could help to classify the goodness of Foursquare photos, and the goodness of these photos could also be another interesting feature to analyze.