CAPSTONE PROJECT - RESTAURANT CLASSIFICATION

NIMA AMINI

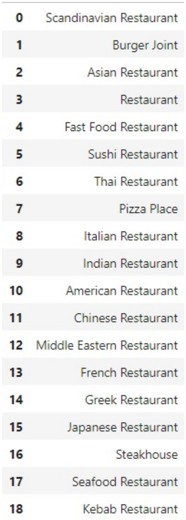
Introduction

Being somewhat disappointed after trying out a new restaurant is quite a common experience in Stockholm county. It would therefore be beneficial for restaurant customers to get some guidance as to whether a restaurant in Stockholm county is worth trying based on a few immediately visible attributes which are easy to input into the model by the customer. Thus our question is:

Can we classify whether a food place in Stockholm county is "good enough" or "not good enough" before ordering in it?

Of course we need to define what "good enough" means. A natural approach is to put a threshold based on the average rating given by customers over all restaurants in Stockholm county. We shall define a restaurant to be "good enough" if its rating is above average and "not good enough" otherwise. Since putting a hard threshold is likely to lead to marginal misclassifications we are also interested in the relative probabilities between "good enough" and "not good enough" to get a sense of how strong the classification is. From our analysis it will fall out that one-hot encoding all existing venue categories and all cities within Stockholm county leads to an inaccurate model due to too many features and not enough data in certain locations. We find that using the the following inputs lead to a model with decent accuracy:

(1) **Venue category**:
A one-hot encoded feature marking exactly one of the following restaurant types (top 19 most common venue categories):



0	Scandinavian Restaurant
1	Burger Joint
2	Asian Restaurant
3	Restaurant
4	Fast Food Restaurant
5	Sushi Restaurant
6	Thai Restaurant
7	Pizza Place
8	Italian Restaurant
9	Indian Restaurant
10	American Restaurant
11	Chinese Restaurant
12	Middle Eastern Restaurant
13	French Restaurant
14	Greek Restaurant
15	Japanese Restaurant
16	Steakhouse
17	Seafood Restaurant
18	Kebab Restaurant

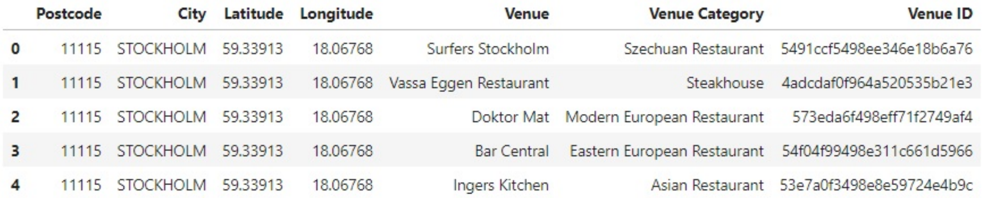
(2) **Price Tier**:
An integer in the range 1-4.

(3) **In Stockholm City**:
A binary variable indicating whether the restaurant belongs to Stockholm city or not.

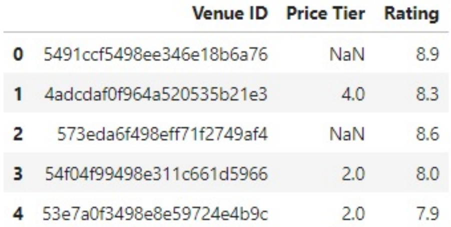
The output will consist of a classification ”good enough” with probability p and ”not good enough” with probability $1 - p$. Thus our aim is to see how much three visible attributes: *venue category*, *price tier* and *location* can tell us about the deviation of restaurant rating from the mean.

Data

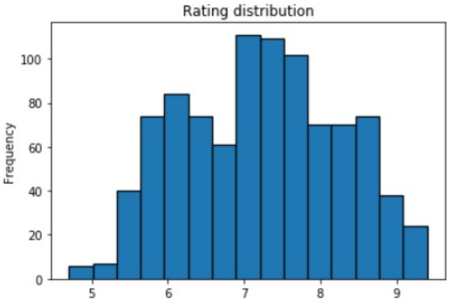
To build our model we first require the geocoordinates of all restaurants in S-tockholm county. One way to collect this data is to look at a small (e.g 500 meter) radius around the geolocation of each postcode in Stockholm county. A list of all postcodes in Stockholm county can be retrieved via the postcode population spreadhsheet provided by SCB (the Swedish government agency responsible for producing official statistics regarding Sweden). To get the geolocation for each postcode we use the ’Here API’. Moreover to explore the venues around each postcode we use the ’Foursquare API’. We filter the results on venue category for words like ’Restaurant’, ’Burger’, ’Food’ and ’Steak’. To each food place venue we retrieve rating (a float 1-9) and price tier (an integer 1-4) via another (premium) call to the ’Foursquare API’.



	Postcode	City	Latitude	Longitude	Venue	Venue Category	Venue ID
0	11115	STOCKHOLM	59.33913	18.06768	Surfers Stockholm	Szechuan Restaurant	5491ccf5498ee346e18b6a76
1	11115	STOCKHOLM	59.33913	18.06768	Vassa Eggen Restaurant	Steakhouse	4adcdaf0f964a520535b21e3
2	11115	STOCKHOLM	59.33913	18.06768	Doktor Mat	Modern European Restaurant	573eda6f498eff71f2749af4
3	11115	STOCKHOLM	59.33913	18.06768	Bar Central	Eastern European Restaurant	54f04f99498e311c661d5966
4	11115	STOCKHOLM	59.33913	18.06768	Ingers Kitchen	Asian Restaurant	53e7a0f3498e8e59724e4b9c



	Venue ID	Price Tier	Rating
0	5491ccf5498ee346e18b6a76	NaN	8.9
1	4adcdaf0f964a520535b21e3	4.0	8.3
2	573eda6f498eff71f2749af4	NaN	8.6
3	54f04f99498e311c661d5966	2.0	8.0
4	53e7a0f3498e8e59724e4b9c	2.0	7.9


Rating distribution
Frequency