CAPSTONE PROJECT - RESTAURANT CLASSIFICATION

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 new 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 "satisfactory" or "not satisfactory" before ordering in it?

We clearly need to be more precise by what we mean by "satisfactory". 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 "satisfactory" if its rating is above average and "not satisfactory" otherwise. Since putting a hard threshold is likely to lead to marginal misclassifications we are also interested in the relative probabilities between the classes to get a sense of how strong the classification is.

What type of restaurant, in which price tier, and in which location maximizes the probability of being a "satisfactory" restaurant?

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 because of too many features and not enough data in certain locations. We find that using the following inputs lead to a better model with decent precision:

(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.

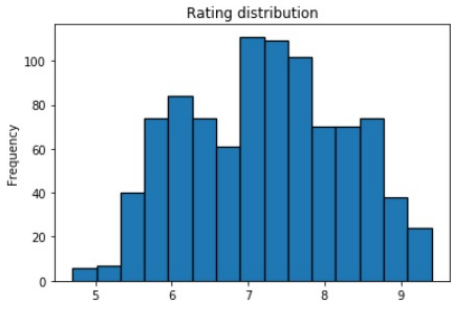
The output will consist of a classification "satisfactory" with probability p and "not satisfactory" with probability $1-p$. Thus our aim is to see how much three visible attributes: *venue category*, *price tier* and *location* can tell us about the deviation of restaurant rating from its mean.

Data

To build our model we first require the geocoordinates of all restaurants in Stockholm 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. We do this to capture restaurants within the boundary of each city in Stockholm county more accurately. 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



Methodology

Since the classification "satisfactory" vs "not satisfactory" will be treated as a fuzzy binary classifier it seems most appropriate to use logistic regression to carry out the classification. This will also produce the required probabilities we require so we can get a sense of how strong the classification is.

In order to prepare the data we have thrown away all restaurants that have no rating. Restaurant that have a rating but no price tier, we have decided to keep and remedy this by assigning it the average price tier of all restaurants which is computed to be 2 (rounded up to nearest integer from ~ 1.97).

The average rating to be used as threshold for our classification is computed to be 7.164276729559743.

We expect that venue category, price tier and location of the restaurant will be able to explain a good portion of the rating of a restaurant, so these will be our preliminary inputs. Note that these attributes are easily observed by the user which is a requirement for our model.

We have 45 cities/towns in Stockholm county that have at least one venue with non-null rating and 62 venue categories. Using one-hot encoding and price tier as a feature this gives us 108 features in total with 795 observable restaurant ratings. Our exploratory analysis shows that using all location data tends to overfit the model and thus not perform well on the test data (roughly 50% precision). To reduce the number of features we pick out only the top 19 venues since these have at least 10 data points. Moreover if we analyze the average rating per city we get the table below.

	City	Mean Rating
0	NACKA STRAND	7.500000
1	DANDERYD	7.500000
2	SIGTUNA	7.450000
3	STOCKHOLM	7.427863
4	ÅRSTA	7.400000
5	DJURSHOLM	7.200000
6	ENSKEDEDALEN	7.142857
7	ÅKERSBERGA	7.100000
8	VALLENTUNA	7.100000
9	SPÅNGA	7.100000
10	SKARPNÄCK	7.100000
11	VÄLLINGBY	7.033333
12	HÄGERSTEN	6.930000
13	ÄLVSJÖ	6.871429
14	BROMMA	6.863158
15	SUNDBYBERG	6.822222
16	GRINDA	6.800000
17	SOLLENTUNA	6.766667
18	SOLNA	6.757500
19	LIDINGÖ	6.630000
20	SÖDERTÄLJE	6.625000
21	TYRESÖ	6.600000
22	VAXHOLM	6.600000
23	MÄRSTA	6.566667
24	FARSTA	6.533333
25	NYNÄSHAMN	6.533333
26	STOCKSUND	6.500000
27	JOHANNESHOV	6.472727
28	KISTA	6.460000
29	TUMBA	6.450000
30	NACKA	6.370000
31	SALTSJÖBADEN	6.300000
32	NORSBORG	6.200000
33	JÄRFÄLLA	6.160000
34	HUDDINGE	6.125000
35	TÄBY	6.122222
36	VÄSTERHANINGE	6.100000
37	ENSKEDE	6.050000
38	SKÄRHOLMEN	6.000000
39	UPPLANDS VÄSBY	5.900000
40	HANDEN	5.750000
41	SALTSJÖ-BOO	5.500000
42	VÅRBY	5.450000
43	SKÖNDAL	5.350000
44	SÖDERBY	5.200000

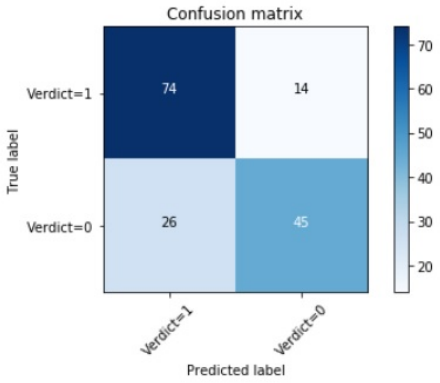
The average ratings range from 7.5 in Nacka Strand down to 5.2 for Söderby at the bottom. However the number of observations for the cities above Stockholm is very limited so we cannot infer any accurate conclusions regarding the correlation between these locations and the general quality of their restaurants. We see however that Stockholm city has a very high average rating which makes intuitive sense since it is much harder for restaurants to make it in the city of Stockholm due to fierce competition and high rents, so we may have some survival bias there. Since a majority of the restaurants belong to Stockholm city in our data set it makes sense to introduce a binary feature that differentiates between whether or not a restaurant lies in Stockholm city as this is likely to affect the rating based on the table above.

This reduces the number of features down to 21.

To perform the logistic regression we first standardize the feature set to a normal distribution with mean 0 and variance 1. We then divide the data into a training set and a test set according to a 80%/20% split with rows chosen at random. Finally we build the logistic regression model using the Python sklearn library on the training set.

Results

Evaluating our model on the test set, our confusion matrix looks as follows:



Here Verdict = 1 corresponds to the positive output "satisfactory" and Verdict = 0 corresponds to a negative output "not satisfactory". We see that the proportion of false positives (25.5%) roughly equals the ratio of false negatives (23.7%) so there is a low bias in our misclassifications. Our overall precision is 75% which is a satisfactory, but not excellent accuracy. Thus our features tend to capture a good portion of the information in the rating. The relative probabilities in the classification for the first 20 restaurants in the test set is given by the below table.

	not good enough	good enough
0	0.180976	0.819024
1	0.553629	0.446371
2	0.487636	0.512364
3	0.286802	0.713198
4	0.382519	0.617481
5	0.327545	0.672455
6	0.327545	0.672455
7	0.570597	0.429403
8	0.487636	0.512364
9	0.365666	0.634334
10	0.407304	0.592696
11	0.824365	0.175635
12	0.407304	0.592696
13	0.553629	0.446371
14	0.180976	0.819024
15	0.407304	0.592696
16	0.365666	0.634334
17	0.658013	0.341987
18	0.254800	0.745200
19	0.407304	0.592696
20	0.377216	0.622784

A summary of the different evaluation metrics are given as follows:

- **Precision**: 75%
- **f1-score**: 0.74
- **Recall**: 0.75

- **Jaccard similarity score**: 0.748
- **LogLoss**: 0.6

Finally running our model over all possible inputs we find that

An **expensive Scandinavian Restaurant in Stockholm City** has the maximum probability ($\sim 80\%$) of being a "satisfactory" restaurant according to customers.

A **cheap Fast Food Restaurant Restaurant not in Stockholm City** has the minimum probability ($\sim 50\%$) of being a "satisfactory" restaurant according to customers.

Discussion

As expected we see that many classifications are a very close call, so we would not expect extremely high accuracy in the range $85 - 100\%$. The relationship between rating and price tier is not linear which is why we decided against using a multilinear regression to predict the actual rating. On one hand a high price is a negative attribute which can highly penalize the rating in case the restaurant does not live up to what it charges. On the other hand as a rule of thumb you should "get what you pay for" so price should be positively correlated with higher rating for restaurants that do their job. Thus price tier must be used together with other features to somehow distinguish this. This is one reason why we would not expect extreme precision in the classification, since it may depend on attributed not immediately visible to customers.

It is perhaps not entirely surprising that Swedish people favour their own cuisine and that such restaurants hold a high standard since owners and chefs have cultural and culinary expertise in the area. Also the best Scandinavian restaurants are likely found in the crowdy areas around Stockholm city (and are therefore more likely to be expensive) which is to be expected. The ratings are perhaps further inflated by the large number of tourists that visit Stockholm city each year and likely have a positive bias towards Scandinavian culture and therefore tend to give higher ratings on Scandinavian restaurants while having less problem spending money since they are on vacation. The recommendation is therefore to build an expensive Scandinavian restaurant in Stockholm city in order to maximize the chance of customer satisfaction.

Finally it is neither surprising that cheap Fast Food Restaurants are least likely to be satisfactory, since the food is of lower quality and you get what you pay for. Moreover fast food restaurants around the city are usually better staffed, have more experienced workers with higher pay, and that work more effectively due to higher demands. Therefore it perhaps makes sense that fast food restaurants inside Stockholm city have a higher chance of being "satisfactory" than those outside Stockholm city. The recommendation is therefore not to build a cheap fast food restaurant outside Stockholm city if customer satisfaction is a priority.

Conclusion

In this report we have built a model based on logistic regression to determine the probability of a restaurant being "satisfactory" vs "not satisfactory" based on attributes: venue category, price tier and location. We found that expensive

Scandinavian restaurants in Stockholm city have the highest chance of being satisfactory whereas cheap fast food restaurants outside Stockholm city have the least probability of being satisfactory. The classification precision is 75%.