Report on hotel ratings in « Ville de Paris »

Péter ENDES-NAGY

Introduction

This is a report on *hotel ratings in City of Paris.* I investigated on the sample of Parisian hotels what factors influence the probability that a hotel is highly rated by their guests.

Data

I used the `hotels-europe` dataset that includes information on features of hotels in 46 European cities and for 10 different dates. A detailed description of the dataset is available here: https://gabors-data-analysis.com/datasets/#hotels-europe

I was interested in factors that influence hotel ratings, like distance and stars, therefore the `hotels-europe_features` table was used. As a first step, I filtered for actual Paris (no banlieux), Hotels only. Missing stars and distance data was also filetered out. A new `highly_rated` dummy variable was constructed that takes the value of 1 if rating is greater than 4, otherwise 0. This `highly_rated` is my dependent variable.

There are 1332 Parisian hotels in the dataset, 56.5% of them are highly rated. Further descriptive statics on the key variables are available in `Table 1`.

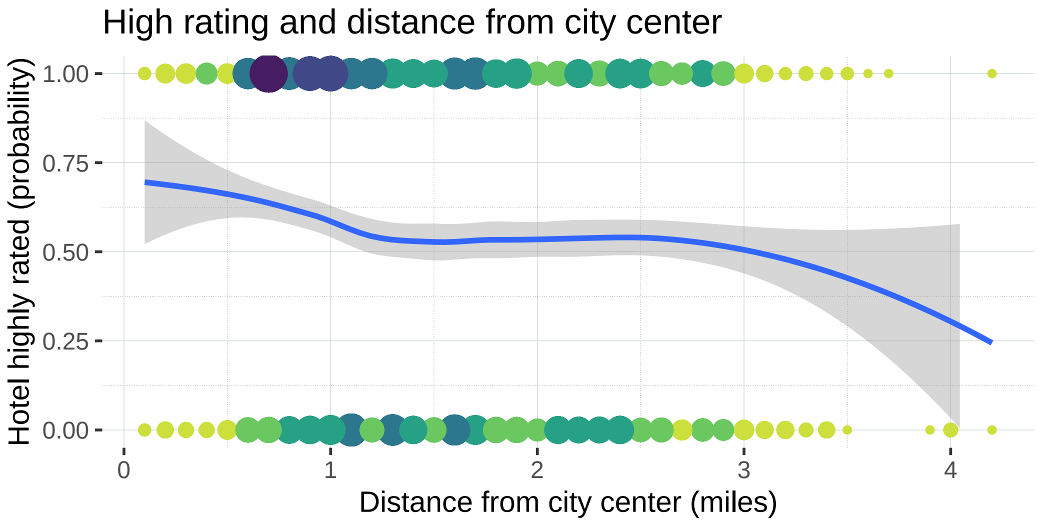
Table 1: Descriptive statstics of Parisian hotels

	Mean	SD	Median	Min	Max	Range	P05	P95	N
Highly rated	0.56	0.50	1.00	0.00	1.00	1.00	0.00	1.00	1332
Distance (miles)	1.61	0.78	1.50	0.10	4.20	4.10	0.60	2.90	1332
Hotel stars	3.22	0.80	3.00	1.00	5.00	4.00	2.00	5.00	1332

data from: https://osf.io/utwjs

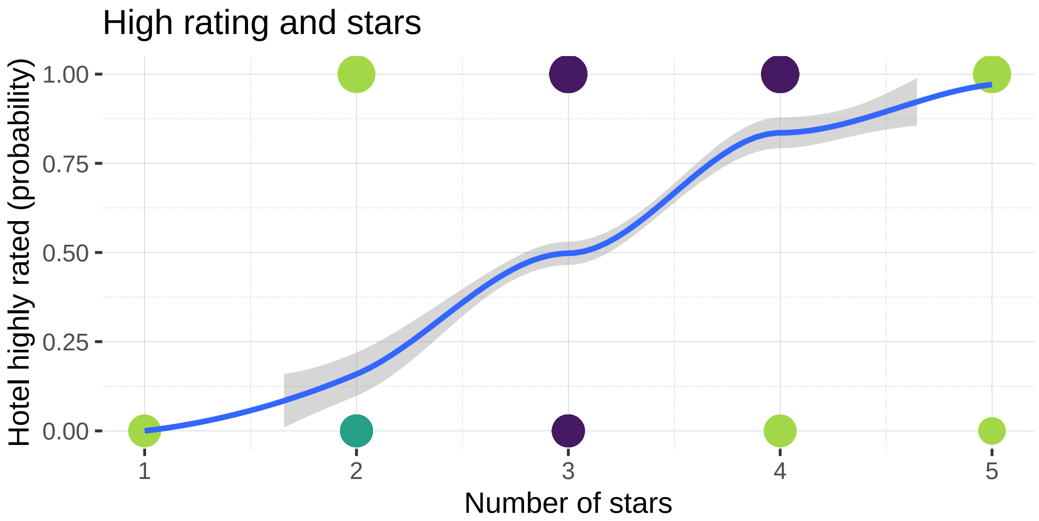
Before building the models, I investigated the main characteristics of the relationships between high rating and the independent variables: distance and stars.

As the Figure 1 shows, the relationship between high rating and distance isn't linear, therefore using `lspline` is recommended with breakpoints at 1.25 and 2.75.



Parisian hotels, data from: https://osf.io/utwjs

Hotel stars are rather ordinal variables than interval/ratio, since the distance between 2 and 3 stars isn't necessary the same as between 3 and 4. Plotting `highly_rated` against hotel stars implies that we can ease the strictness and use it as continuous variable. The relationship seems more or less linear.



Parisian hotels, data from: https://osf.io/utwjs

Modelling likelihood of high rating

For modeling the relationship between distance, stars and high rating, I run a simple LPM, a logit and probit regression. I also included results with and without using splines for distance.

The regression results are available in `Table 2`:

The coefficients (marginal in case of logit and probit since the raw coefficients can't be interpreted) are basically the same, so there is no difference in interpreting them. Each additional star means 30% points higher likeliness that a Parisian hotel has a high rating. Without use of splines, distance is significant on 5% only, a hotel a mile further away from the center is 3% points less likely to have a high rating. With splines,

Table 2: Probability of having a high rating among Parisian Hotels : Model summaries

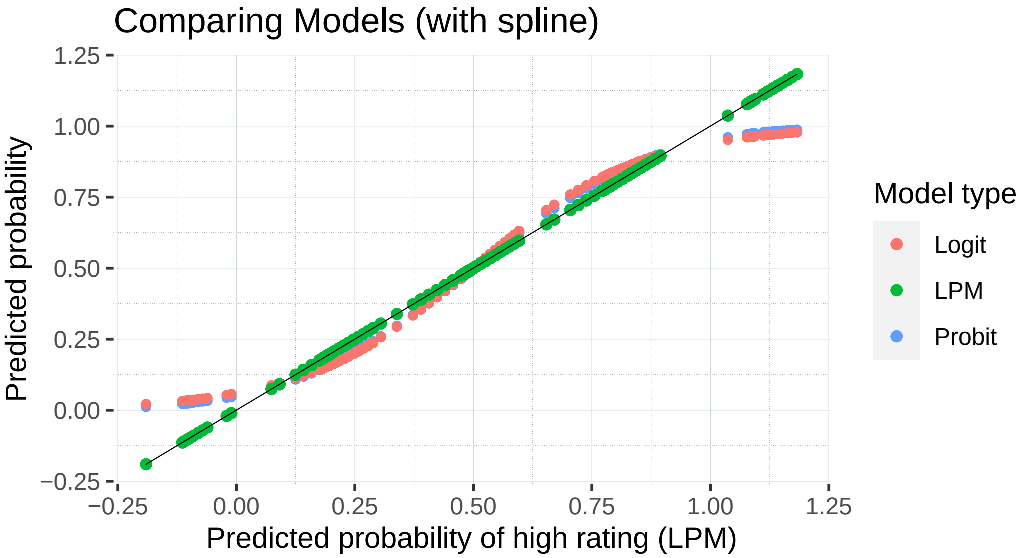
	Linear	Linear (spline)	Logit	Logit (spline)	Probit	Probit (spline)
stars	0.2995**	0.2983**	0.3074**	0.3059**	0.3045**	0.3031**
	(0.0150)	(0.0150)	(0.0277)	(0.0276)	(0.0119)	(0.0120)
distance	−0.0302*		−0.0288		−0.0306*	
	(0.0153)		(0.0154)		(0.0152)	
Distance < 1.25		−0.1016		−0.1041		−0.1004
		(0.0561)		(0.0577)		(0.0566)
Distance 1.25-2.75		0.0131		0.0143		0.0101
		(0.0287)		(0.0280)		(0.0281)
Distance > 2.75		−0.1691		−0.1719		−0.1703
		(0.1017)		(0.1071)		(0.1037)
R2	0.237	0.239				

data from: https://osf.io/utwjs

* p < 0.05, ** p < 0.01

the distance doesn't seem to matter, none of the coefficients are significant at 5% - hotels with more stars are probably concentrated in the city center.

The difference between LPM, logit and probit models lie in the predicted values. Logit and probit keeps the predicted probabilities between 0 and 1, in our models [0.01 , 0.99] and [0.02 , 0.98] for probit and logit respectively, while in the LPM model, they range between -0.19 and 1.18.



Conclusion

In this report, I investigated how distance from the city center and number of stars influence the likely-hood that a Parisian hotel has high rating, above 4. Based on my dataset, hotels with more stars are significantly more likely to have a high rating. Distance is insignificant in my model.

Generalizing the results on other cities shall be carried out carefully. Cities have different structures, the role of distance might vary greatly from one city to another - "better" hotels might be concentrated in different belts or areas. The insights on hotel stars might be generalized as they measure the quality and comfort level of given hotels based on international standards.