DA2 Assignment 2 - Barcelona Hotel Ratings

Data

From the Hotels-Europe dataset I analyzed Hotels in Barcelona, how different variables relate to the probability of being highly rated by customers. High rating means 4+.

I included ratings from 2017 December weekdays, where key variables were not missing. A single hotel was priced above 4500 Euros. The hotel had 1 star and was nearly twice as expensive as the second one. I considered it to be an erroneous record and excluded it. Lastly, I created ln price variable since the price distribution was lognormal and the main binary variable which is 1 for highly rated hotels and 0 for ratings below 4.

Data Summary

Table 1: Descriptive statistics

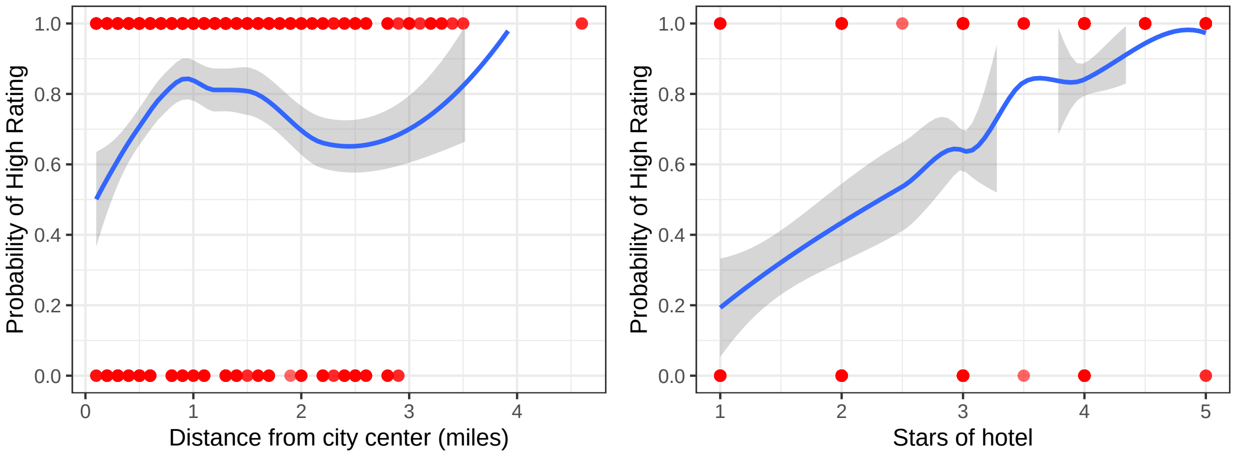
	Mean	Median	SD	Min	Max	P05	P95	N
High or Low Rating (1/0)	0.74	1.00	0.44	0.00	1.00	0.00	1.00	674
Avg. rating	4.10	4.10	0.40	2.20	5.00	3.50	4.60	674
Price	352.17	288.00	305.94	52.00	2580.00	71.65	861.70	674
Ln Price	5.50	5.66	0.88	3.95	7.86	4.27	6.76	674
Distance from center	1.19	1.00	0.80	0.10	4.60	0.20	2.80	674
Stars	3.51	4.00	0.96	1.00	5.00	2.00	5.00	674
Number of nights	2.46	1.00	1.50	1.00	4.00	1.00	4.00	674

The table shows variables I considered to include in the models. Missing is the tripadvisor rating which correlate with user ratings. In my opinion this should be used for external validity rather than as a control. The lowess graph of Number of nights with High rating showed no association: same mean and spread of ratings regardless of customers spending 1 or 4 nights, thus I excluded it. Important findings in the table:

- Most, 71%, of hotels are highly rated.Number of observations is 674.
- There are some huge extreme values in prices and distance.

Regression exploration and findings

The 3 explanatory variables are distance, stars and log prices. Below are the loess regressions of distance and stars.



Distance has an interesting pattern. For hotel less than 1 mile from the center, distance is positively related with the chance of high rating. Between 1 and 2 miles the opposite is true. For longer distances farther hotels are more likely to be positively rated. Possible explanations include: more crowded areas are less enjoyable. The city might have multiple centers, so hotels far from this point might be closer to other attractions.

Stars are straight forward. Hotels with more stars generally have higher chance of good ratings.

I included distance with 2 splines and stars with a linear regression in my models. I found the graph of log prices to have 2 knots. See the log price and other regressions in the Appendix.

Regression Models

Model 1 shows the **linear probability model**. A hotel with 1 extra star has 14.6% higher chance for good rating with the same distance from city center and price level. The standard error is relatively small, this is a significant finding. The difference between a hotel 0 and 1 mile from the city center is 16.4% higher chance for good rating with stars and prices being equal. This finding is less practical since no hotel is exactly at the city center. Any distance father the differences are not significant. Even in the 0-1 mile range, the standard error is high, the difference is only significant at 95% confidence level. As for prices, under the first spline (ln price = 5) we find a positive association, 1% higher price is associated with 0.5% higher chance for good rating with distance and stars being the same. Between the first and second knot we find nearly the opposite while above the second spline (ln price = 5.6) a smaller positive association. The log price levels are hard to interpret. They roughly translate back to 150 and 270 Euros. At all 3 intervals the findings are significant.

Model 2 shows the **coefficients of the logit model**. The logit model transforms the LPM model so that no predictions can be outside the 0-1 range. We don't interpret its coefficients.

Rather **Model 3** shows the **marginal effect** of the **logit** model. These have the same interpretation as the LPM coefficients. We can see they are slightly lower with the same level of significance.

Models 4 and 5 show the coefficients then the marginal effects of the **probit** model. Slightly below the LPM the marginal effects have the same explanation.

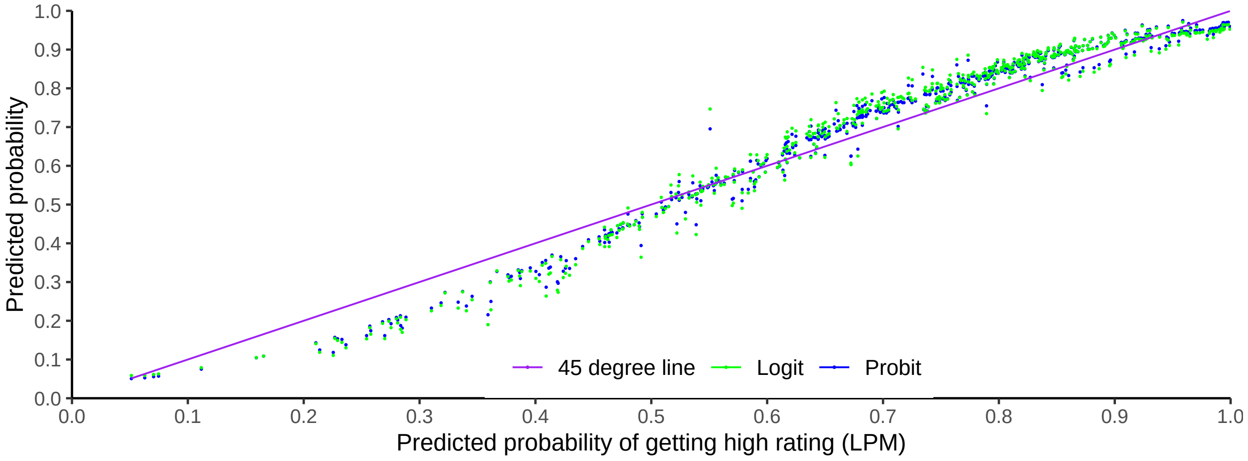
The model differences are not statistically significant as the confidence intervals heavily overlap. Hotels with high chance at a good rating, tend to have 5 stars, be about 1 mile away from the center and have higher prices except in the second quarter (150 to 270 Euros) of prices.

	Model 1	Model 2	Model 3	Model 4	Model 5
Constant	−2.222**	−17.649**		−10.125**	
	(0.427)	(3.145)		(1.756)	
stars	0.146**	0.838**	0.121**	0.501**	0.125**
	(0.018)	(0.120)	(0.021)	(0.069)	(0.015)
Distance under 1 mile	0.164*	1.067*	0.153*	0.603*	0.151*
	(0.075)	(0.461)	(0.068)	(0.268)	(0.066)
Distance 1 to 2 miles	−0.059	−0.296	−0.043	−0.188	−0.047
	(0.060)	(0.412)	(0.059)	(0.235)	(0.059)
Distance above 2 miles	0.014	0.049	0.007	0.045	0.011
	(0.050)	(0.456)	(0.066)	(0.256)	(0.064)
Log price below 1st spline	0.506**	3.323**	0.478**	1.896**	0.474**
	(0.096)	(0.696)	(0.111)	(0.390)	(0.093)
Log price 1st to 2nd spline	−0.543**	−4.396**	−0.632**	−2.418**	−0.605**
	(0.102)	(0.863)	(0.139)	(0.480)	(0.115)
Log price above 2nd spline	0.190**	2.030**	0.292**	1.079**	0.270**
	(0.050)	(0.532)	(0.082)	(0.292)	(0.071)
Num.Obs.	674	674	674	674	674

* $p < 0.05$, ** $p < 0.01$

Comparing model predictions

Below is a comparison between the predictions of the models. They differ more as we go towards 0 and 1. In the Appendix I show that only the LPM has predictions above 1.


Predicted probability
45 degree line
Logit
Probit
Predicted probability of getting high rating (LPM)