DA2-Assignment-2

Gyongyver Kamenar (2103380)

Introduction In this assignment I analyzed the relationship between hotel rating, distance from the city centre and stars of the hotel in Rome. The used dataset with hotels in Europe was provided by the textbook and the hotels in Rome were selected. I built 4 different models to predict the probability that the hotel is highly rated (at least 4) based on the distance and the number of stars.

Data exploration The datatable contains accomodations in Europe, so I filtered for Rome and then I explored the accomodation types. I found several different types of accomodation in the datatset but I decided to narrow down the research question just to hotels. the analysis will better represent the hotels in Rome. Besides, I omitted observation with missing distance, rating or stars and also few observations with not integer stars.

To explore the data I checked the descriptive statistics (Table 1) of the most relevant numerical variables and also some dummy variables like offer, holiday, weekend and the highly rated, which is our variable of interest. Furthermore, I examined the correlation between variables as illustrated in the first plot. The variable of my interest (highly_rated) is obviously highly correlated with rating, but also with stars and price, and there is significant but negative correlation with distance too. Finally, I plotted rating with distance and stars with the loess method, which suggest that splines on distance might be useful.

Models and interpretation The first model is the linear probability model with distance and and stars as explanatory variables and highly rated as the dependent variable. The huge drawback of this model is that the estimeted probability is not limited so practically it can be below zero or above 1 as illustrated on Figure 2. The estimated coefficients are presented in Table 2. Both explanatory variables are significant at 1% level. The coefficient of distance means, that with the same stars hotels 1 unit further from the centre are 0.036 less likely to be highly rated on average. Considering the stars coeffeecient,among hotels with the same distance, 1 more star hotels are 0.16 more likely to be highly rated on average.

The second model used the same variables but it is a logit model. Because of the functional formula the predicted probability is limited between 0 and 1. I calculated the average marginal differences (also included in Table 2) which can be interpreted similarly to the LPM. Based on the logit model, on average hotels 1 unit further are 0.045 less likely and hotels with 1 more start are 0.18 more likely to have high rating ceteris paribus.

The third one is a probit model which is very similar to the logit. The functional form is also limits the values at 0 and 1, the estimated average marginal differences for distance and starts are -0.040 and 0.176 which are between the LPM and logit but really close to both, practically it has just minor difference.

Additionally, I estimated a logit model with splines on distance because of the suggestion at the EDA as determined the breakpoint by build-in function. In this model, there is a coefficient for hotels with distance<=1.2 and and one for distance>1.2. Thus, the marginal coefficients can be interpreted as for ususal logit, but with a distinction between the distance categories.

Prediction I calculated the predicted probabilities based on the 3 models and illustrated on Figure 2. The 45 degree line helps us to compare with LMP. As we can see, the LPM predictions go below zero in some cases while the logit and probit curves flatten to 0 and 1. However, around the mean the slope (marginal difference) of the 3 models are almost almost the same.

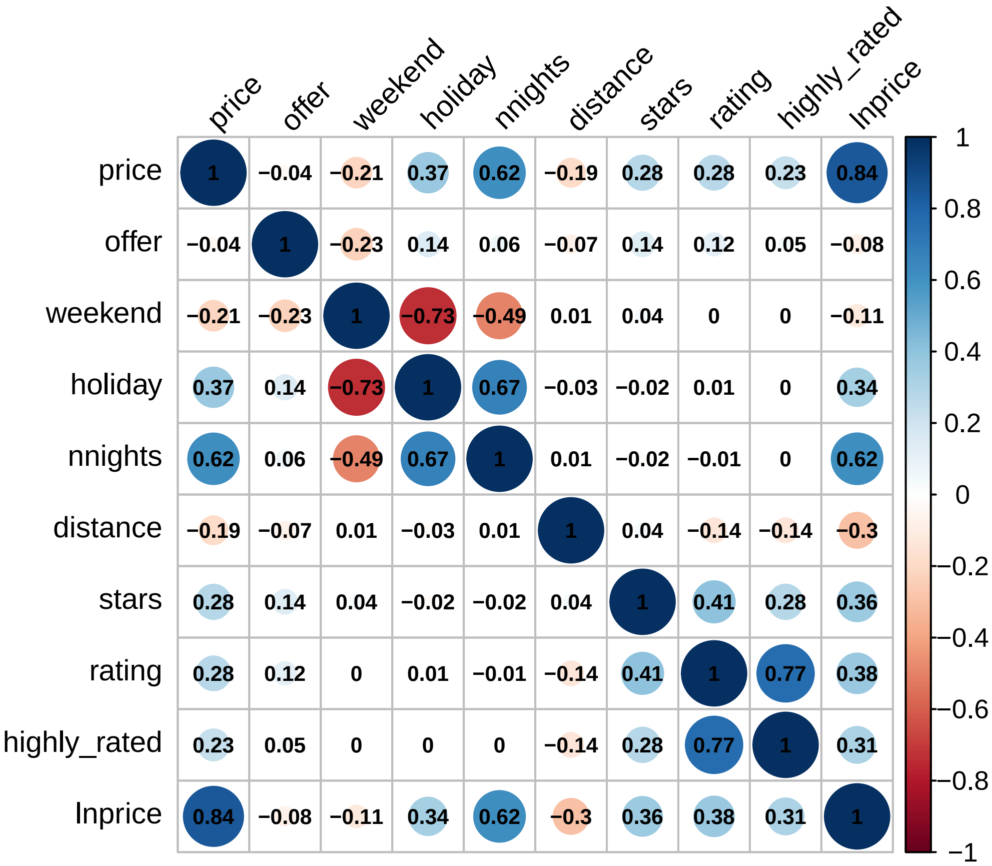
Summary In this paper I used a linear probability model, a logit, a probit and an additional logit model with splines on distance to predict the probability that a hotel is highly rated explained by distance and stars.Both the distance and stars are significant in the models. The marginal differences of the models are almost identical, however logit and probit models have the adventage that they are limited at 0 and 1. Besides, based on the Bayesian information criteria, the logit model with splines is better than any of the others.

Appendix

Table 1: Descriptive statistics

	Mean	Median	SD	Min	Max	P05	P25	P75	P95
price	208.47	139.00	222.93	32.00	4234.00	56.00	90.00	239.00	593.95
distance	1.73	1.10	2.07	0.10	16.00	0.30	0.70	1.70	6.20
offer	0.68	1.00	0.47	0.00	1.00	0.00	0.00	1.00	1.00
weekend	0.61	1.00	0.49	0.00	1.00	0.00	0.00	1.00	1.00
holiday	0.26	0.00	0.44	0.00	1.00	0.00	0.00	1.00	1.00
stars	3.29	3.00	0.91	1.00	5.00	2.00	3.00	4.00	5.00
rating	3.85	3.90	0.50	1.00	5.00	3.00	3.50	4.10	4.50
number of nights	1.41	1.00	1.03	1.00	4.00	1.00	1.00	1.00	4.00
Log(price)	5.03	4.93	0.72	3.47	8.35	4.03	4.50	5.48	6.39
highly rated	0.48	0.00	0.50	0.00	1.00	0.00	0.00	1.00	1.00

Correlation matrix



	price	offer	weekend	holiday	nnights	distance	stars	rating	highly_rated	lnprice
price	1	−0.04	−0.21	0.37	0.62	−0.19	0.28	0.28	0.23	0.84
offer	−0.04	1	−0.23	0.14	0.06	−0.07	0.14	0.12	0.05	−0.08
weekend	−0.21	−0.23	1	−0.73	−0.49	0.01	0.04	0	0	−0.11
holiday	0.37	0.14	−0.73	1	0.67	−0.03	−0.02	0.01	0	0.34
nnights	0.62	0.06	−0.49	0.67	1	0.01	−0.02	−0.01	0	0.62
distance	−0.19	−0.07	0.01	−0.03	0.01	1	0.04	−0.14	−0.14	−0.3
stars	0.28	0.14	0.04	−0.02	−0.02	0.04	1	0.41	0.28	0.36
rating	0.28	0.12	0	0.01	−0.01	−0.14	0.41	1	0.77	0.38
highly_rated	0.23	0.05	0	0	0	−0.14	0.28	0.77	1	0.31
lnprice	0.84	−0.08	−0.11	0.34	0.62	−0.3	0.36	0.38	0.31	1

Figure 1: Exploratory data analysis

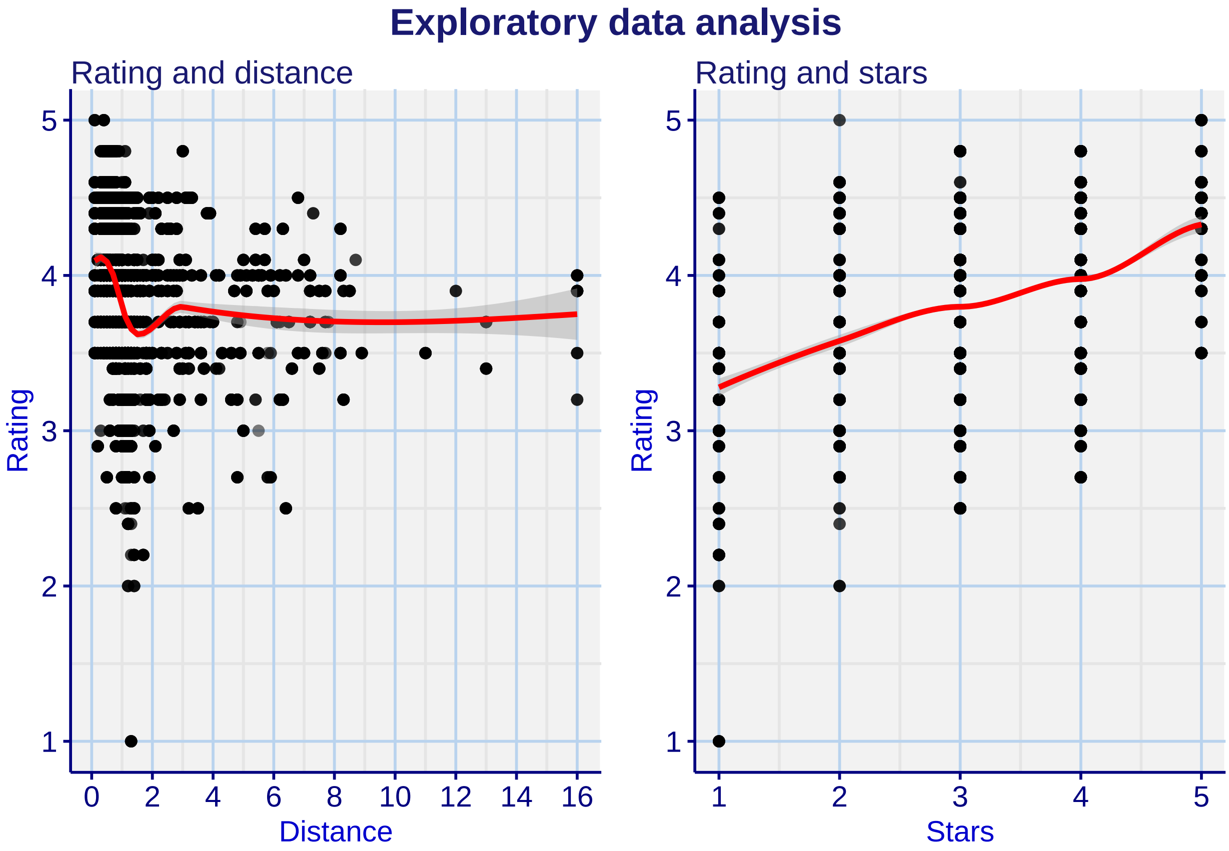
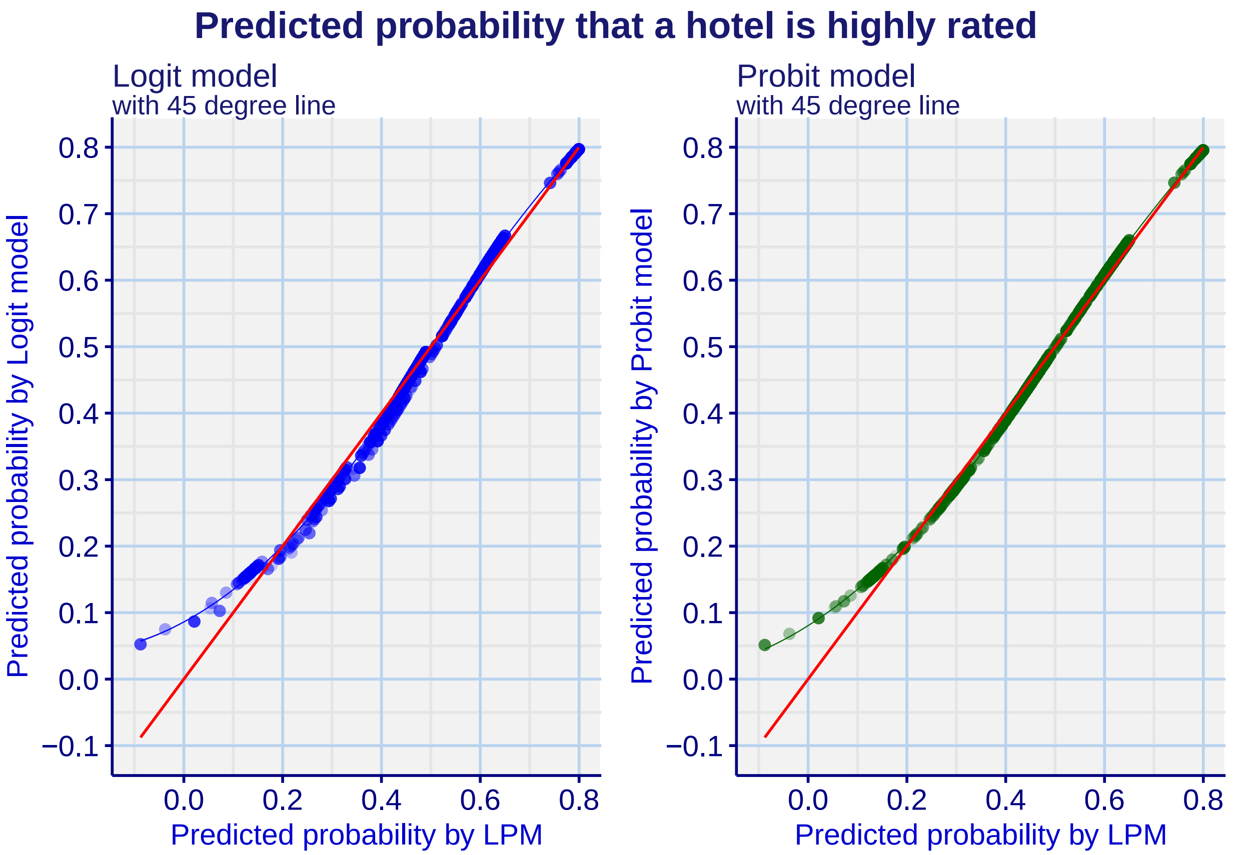


Table 2: Estimated models

	LPM	Logit	Logit marginal	Probit	Probit marginal	Splines logit	Splines marginal
Constant	0.013	−2.195**		−1.349**		−1.124**	
	(0.022)	(0.121)		(0.071)		(0.159)	
distance	−0.036**	−0.180**	−0.045**	−0.101**	−0.040**		
	(0.003)	(0.020)	(0.005)	(0.013)	(0.005)		
stars	0.160**	0.727**	0.181**	0.443**	0.176**	0.684**	0.171**
	(0.006)	(0.035)	(0.009)	(0.021)	(0.008)	(0.036)	(0.009)
lspline(distance, 1.2)1						−1.259**	−0.314**
						(0.104)	(0.026)
lspline(distance, 1.2)2						−0.073**	−0.018**
						(0.019)	(0.005)
Num.Obs.	5182	5182	5182	5182	5182	5182	5182
BIC	6971.1	6622.8	6622.8	6630.0	6630.0	6505.1	6505.1
Std.Errors	Heteroskedasticity-robust	Heteroskedasticity-robust		Heteroskedasticity-robust		Heteroskedasticity-robust	

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

Figure 2: Predicted probabilities


Predicted probability that a hotel is highly rated
Logit model
with 45 degree line
Predicted probability by Logit model
Predicted probability by LPM
Probit model
with 45 degree line
Predicted probability by Probit model
Predicted probability by LPM