Máté Schieszler
Péter Szilvási

Data Analysis 2- Assignment 2

Github: https://github.com/mateschieszler/DA2_2023

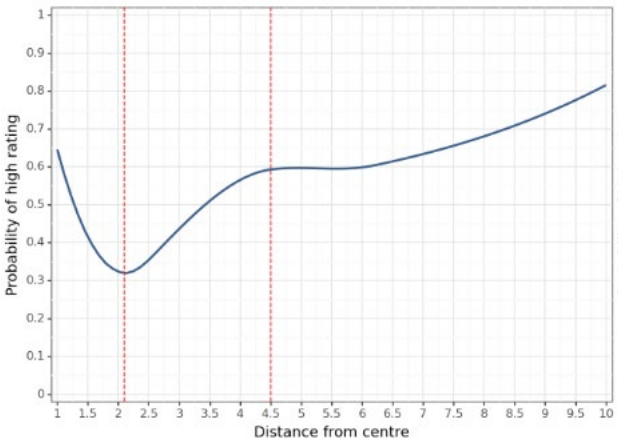
Chosen city: Milan

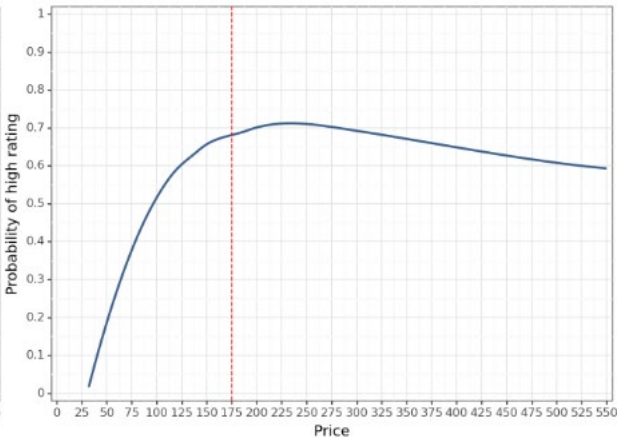
1. Data Cleaning and Filtering:

We dropped null values for rating column as it would lead to a `highly_rated` = 0. We also dropped null values for stars column as we have enough number of observations with 537 accommodations. For prices, to filter out extreme values we dropped values that were beyond the 95th percentile. While for distance we dropped extremes beyond 10kms as we are only looking at Milan, not including smaller cities around it.

2. LPM MODEL & PREDICTION

In our Linear Probability Model we examined how high rating is related to the hotel's distance from the city center and its price. After examining lowess regressions for price and distance we decided to include the explanatory variables as piecewise linear splines with knots at 2.1 and 4.5 for distance and at 175 euros/night for price.





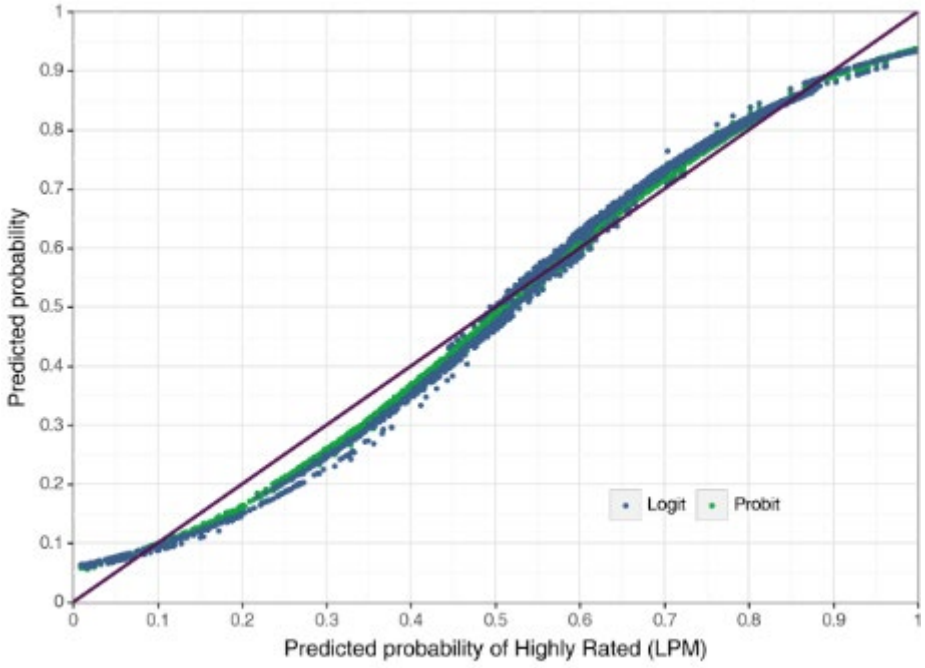
	Dependent variable: highly_rated
	(1)
Stars	0.206***
	(0.009)
Distance less than 2.1	-0.119***
	(0.019)
Distance is between 2.1 and 4.5	0.052***
	(0.013)
Distance more than 4.5	0.036***
	(0.007)
Price less than 175	0.002***
	(0.000)
Price more than 175	-0.001***
	(0.000)
Observations	3401
R^2	0.253
Adjusted R^2	0.252
Residual Std. Error	0.431 (df=3394)
F Statistic	265.273*** (df=6; 3394)
Note:	*p<0.1; **p<0.05; ***p<0.01

This model shows us that in Milan, accommodations with similar features with a higher star rating are significantly 20.6pp more likely to be highly rated, even when controlling for distance from the city center and price category. The probability of a high rating decreases initially with greater distance from the center (-11.9pp) but then reverses (5.2pp and 3.6 after 4.5 kms), indicating a nuanced location-rating relationship. Additionally, while the likelihood of a high rating initially increases with nightly price (2pp/10eur), it starts to decline beyond a certain point by 1pp/10 EUR. Notably, the top 1% of predicted highly rated accommodations in Milan has an average price of 250 EUR/night, emphasizing a non-linear and statistically significant association

between pricing and high ratings. These coefficients are all statistically significant at 1%.

3. Logit & Probit Models

While our logit and probit models are indistinguishable, our coefficients change only slightly compared to our LMP while all remaining statistically significant. However, they produce better predicted values as at tails, probit and logit models are less extreme.



4. Goodness of fit & Conclusion

The logit and probit models exhibit comparable predictive performance, with the logit model showing a slightly better fit. This improvement is attributed to the logit model's ability to capture more distinct probability model means and medians within the y=0 and y=1 groups. Our Brier-score also suggests that logit (0.181) and probit (0.182), both provide a better fit than our LPM (0.186). The log-loss values of each model show the similar ordering: logit (-0.545) and probit (-0.547) are better than LMP (-0.756).