Data Analysis : Assignment 2

Ghazal Ayobi and Shah Ali Gardezi

Introduction

The Question of this case study is how hotels stars is related to highly rated hotels. For this assignment we use Hotels-Europe data. This data set contains two tables **Features** and **Price**. We joined the two tables using left join.

Data Transformation

As a process of filtering and data transformation, we use hotel user rating as the dependent variables and transformed it to a binary variable called *highly_rated* which equals to one if *rating* is more than *4*, *0* otherwise. We selected ***Paris*** City and considered ***Hotels*** as accommodation type. Moreover, we excluded hotels with less than *USD 600* per night, and we removed null and duplicated values from the data set. In order to understand what functional form to include in the regression we examined *Lowess* regression with highly rated hotels and distance. *Figure 1* in the appendix shows that distance and highly rated hotels have negative relationship from city center to 1.2 miles away from center, however *highly_rated* hotels does not indicated any relationship between 1.2 and 3 miles. After 3 miles, as distance changes by one mile highly rated declines. We created a binary variable for *Stars* called *top_stars* which equals to one if *stars* is more than *3*, *0* otherwise. Other variables are *price* which was transformed to log of Price and *weekend* which is binary variable

Analysis

- Filtered the data for **Paris**
- Selected Hotels and Apartments for accommodation type
- Price is less than 600
- Removed null values from *stars*
- Removed duplicates
- Removed Null values from *rating*, *stars*, and *distance*
- Created log of price, *lnprice*
- Created *highly_rated* if rating >= 4

```
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
```

Table 1: Summary Statistics

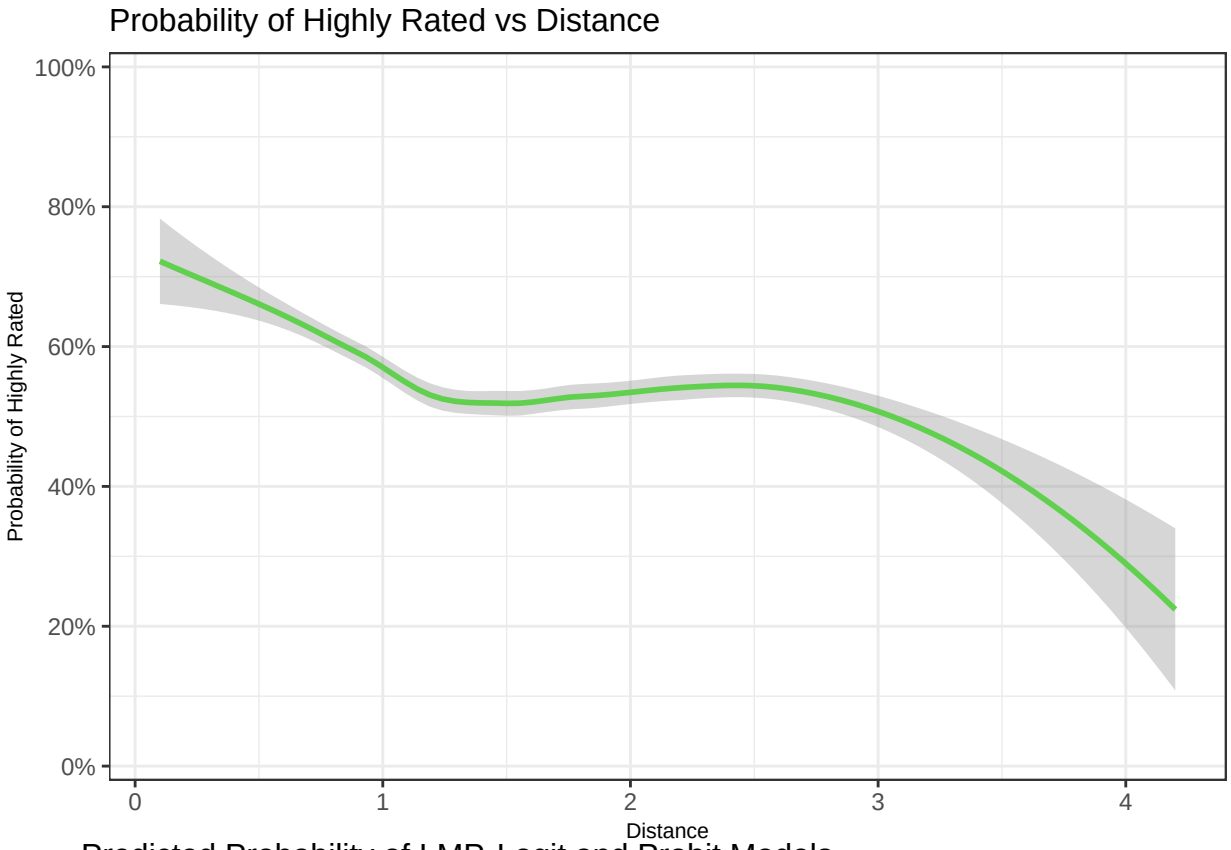
	Mean	SD	Min	Max	Median	P95	N
highly_rated	0.56	0.50	0.00	1.00	1.00	1.00	11397
distance	1.62	0.78	0.10	4.20	1.60	2.90	11397
stars	3.22	0.78	1.00	5.00	3.00	4.00	11397

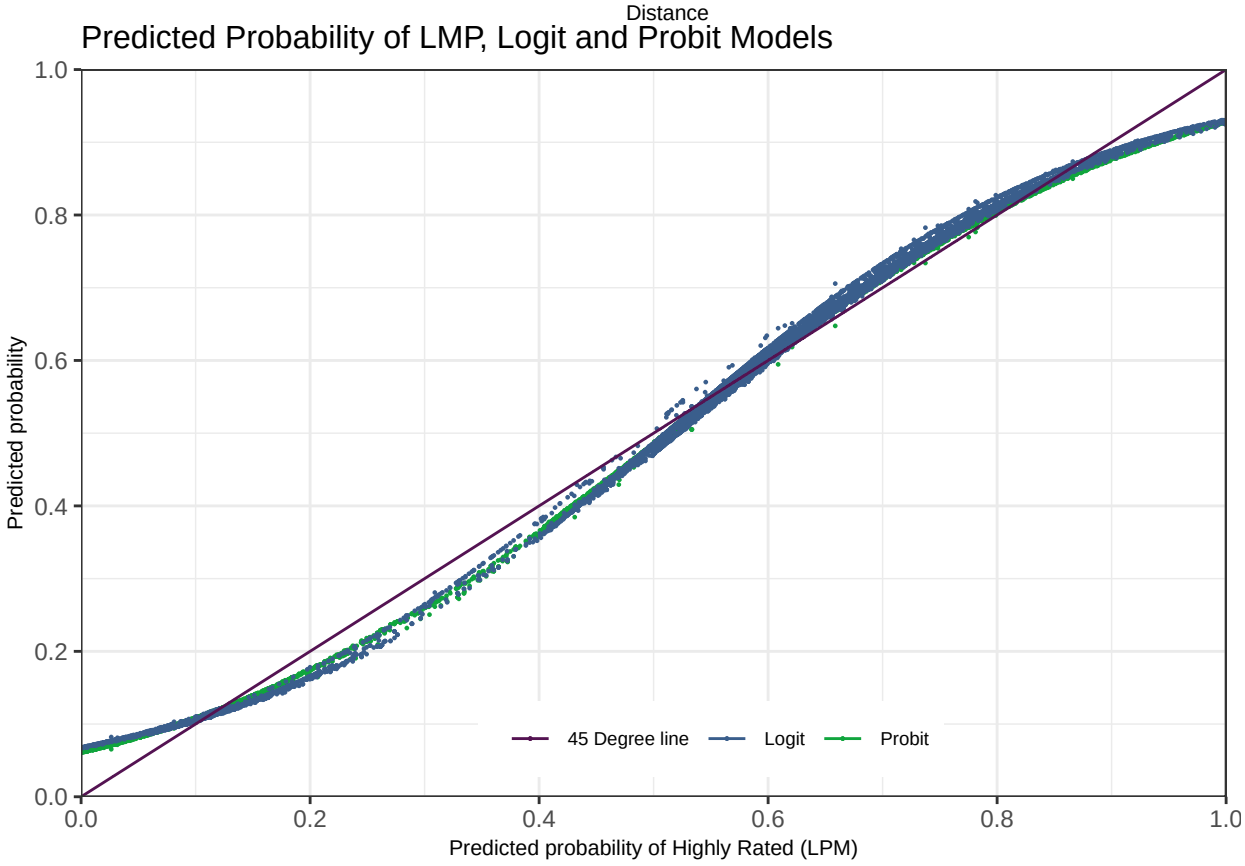
	Model 1	Model 2	Model 3	Model 4	Model 5
Constant	0.027	−2.551**		−1.451**	
	(0.027)	(0.146)		(0.084)	
top_stars	0.421**	2.145**	0.436**	1.266**	0.432**
	(0.012)	(0.075)	(0.012)	(0.041)	(0.011)
lspline(distance, c(1.2, 3))1	−0.108**	−0.518**	−0.104**	−0.340**	−0.113**
	(0.021)	(0.110)	(0.022)	(0.065)	(0.021)
lspline(distance, c(1.2, 3))2	0.018*	0.096*	0.019*	0.061*	0.020*
	(0.009)	(0.043)	(0.009)	(0.026)	(0.009)
lspline(distance, c(1.2, 3))3	−0.251**	−1.268**	−0.254**	−0.802**	−0.268**
	(0.050)	(0.259)	(0.050)	(0.155)	(0.049)
price	0.001**	0.006**	0.001**	0.003**	0.001**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
weekend	0.101**	0.481**	0.098**	0.313**	0.106**
	(0.010)	(0.049)	(0.010)	(0.030)	(0.010)
Num.Obs.	11 397	11 397	11 397	11 397	11 397

* p < 0.05, ** p < 0.01

	Model 1	Model 2	Model 3
Constant	0.027	−2.551**	−1.451**
	(0.027)	(0.146)	(0.084)
top_stars	0.421**	2.145**	1.266**
	(0.012)	(0.075)	(0.041)
lspline(distance, c(1.2, 3))1	−0.108**	−0.518**	−0.340**
	(0.021)	(0.110)	(0.065)
lspline(distance, c(1.2, 3))2	0.018*	0.096*	0.061*
	(0.009)	(0.043)	(0.026)
lspline(distance, c(1.2, 3))3	−0.251**	−1.268**	−0.802**
	(0.050)	(0.259)	(0.155)
price	0.001**	0.006**	0.003**
	(0.000)	(0.000)	(0.000)
weekend	0.101**	0.481**	0.313**
	(0.010)	(0.049)	(0.030)
Num.Obs.	11 397	11 397	11 397
R2	0.185		
PseudoR2		0.149	0.148

* p < 0.05, ** p < 0.01


Probability of Highly Rated vs Distance
100%
80%
60%
40%
20%
0%
Probability of Highly Rated
0
1
2
3
4
Distance


Predicted Probability of LMP, Logit and Probit Models
1.0
0.8
0.6
0.4
0.2
0.0
Predicted probability
45 Degree line
Logit
Probit
0.0
0.2
0.4
0.6
0.8
1.0
Predicted probability of Highly Rated (LPM)