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.

Anaalysis

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

Table 2 shows six regression models: lpm0 lpm, logit, marginal logit, probit and marginal probit. The constant in the model 1, *lmp0* indicates that *13.7%* of low star hotels are highly rated. Top stars hotels are *49.7* percentage point more likely to be highly rated which mean *63.4%* of them are highly rated. The 95% confidence interval around the slope parameter is [0.473, 0.521] which implies that we can be 95% confident that top stars in the sample are highly rated with the higher probability of 47% to 52%. Moreover, we examined lowess regression with distance as the figure 1 shows, we included distance as a piecewise linear spline with knots at 1.2 and 3 miles. In the regressions Table 2, second column, LMP, the coefficient for top_stars is *0.421* comparing hotels with the same distance from city center, price and if it is weekend, the hotles with the top stars are 42.1% more likely to be highly rated. The 95% confidence interval is [0.397, 0.445], and it does not contain zero which indicates that top stars is positively related with highly rated hotels. The other variables also show interesting results.

By looking at the logit and probit estimates for the model, the probability of highly rated to top stars, distance and conditional on price and weekend are same variables as linear model. By looking to the column 3 and 4, the Logit Coefficients are almost five times the size of corresponding logit marginal differences. Furthermore, in the column 5 and 6, probit coefficient is almost three times the size of corresponding probit marginal differences. It is interesting to observe that the two marginal differences, logit and probit, are the same and they are the same with LMP coefficients in column 2 which is applicable for of the independent variables. To generalize the result, it shows that hotels with top stars other things (distance, price and if it is weekend) the same are highly rated. To sum, top stars hotels have a 43 percent points higher chance to be highly rated.

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

Table 2: The Probability of Highly rated hotels and top stars - LMP, Logit, and Probit models

	(1) LMP0	(2)LMP	(3) logit coeffs	(4) logit Marg	(5) Probit	(6) Probit Marg
Constant	0.527**	−0.984**	−7.516**		−4.521**	
	(0.022)	(0.058)	(0.320)		(0.188)	
top_stars	0.431**	0.315**	1.598**	0.325**	0.962**	0.326**
	(0.009)	(0.010)	(0.054)	(0.011)	(0.031)	(0.010)
lspline(distance, c(1.2, 3))1	−0.108**	−0.040	−0.218	−0.041	−0.135*	−0.043*
	(0.021)	(0.021)	(0.112)	(0.021)	(0.067)	(0.021)
lspline(distance, c(1.2, 3))2	0.008	0.022*	0.101*	0.019*	0.055*	0.017*
	(0.009)	(0.009)	(0.045)	(0.009)	(0.027)	(0.009)
lspline(distance, c(1.2, 3))3	−0.241**	−0.243**	−1.272**	−0.240**	−0.742**	−0.234**
	(0.050)	(0.049)	(0.274)	(0.045)	(0.159)	(0.042)
lnprice		0.274**	1.394**	0.263**	0.839**	0.264**
		(0.010)	(0.054)	(0.012)	(0.032)	(0.010)
weekend		0.093**	0.533**	0.100**	0.316**	0.099**
		(0.010)	(0.051)	(0.010)	(0.030)	(0.010)
Num.Obs.	11 397	11 397	11 397	11 397	11 397	11 397

* p < 0.05, ** p < 0.01

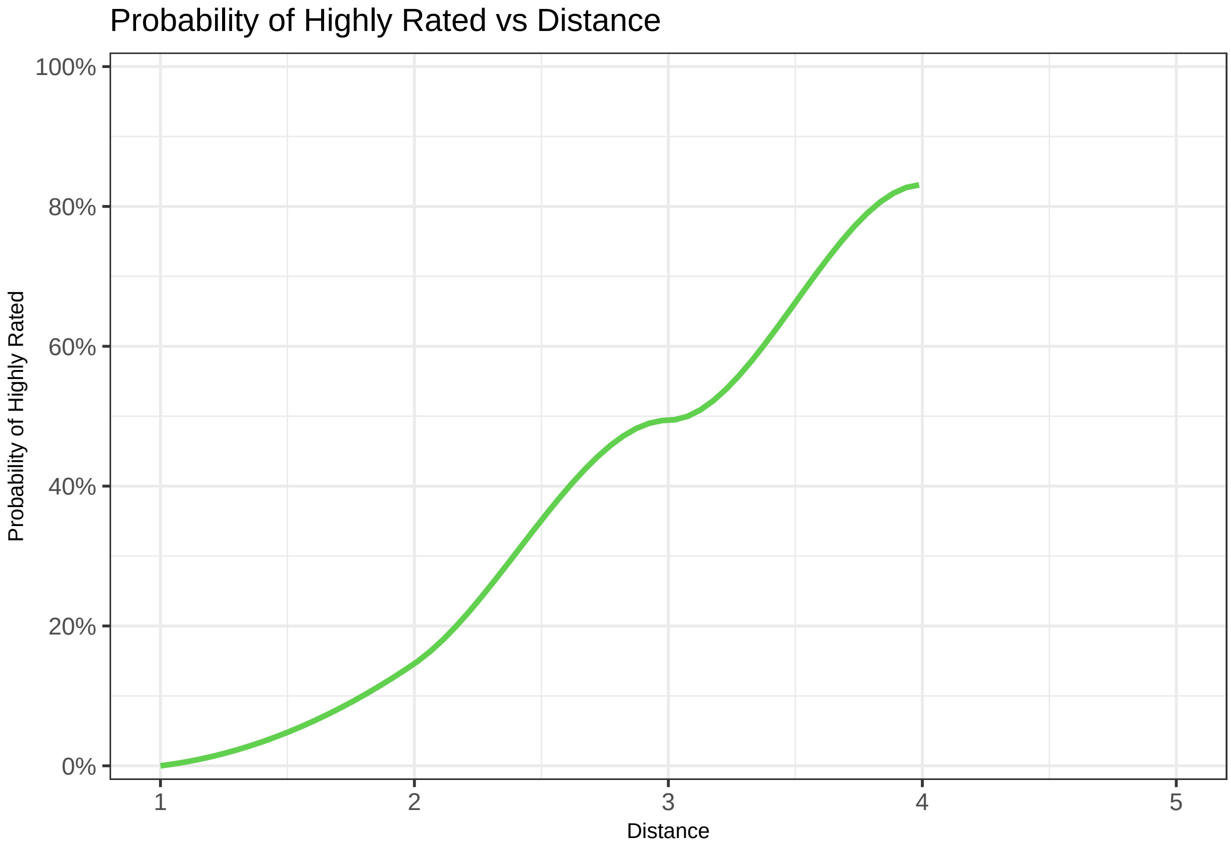
	Model 1	Model 2	Model 3
Constant	−0.984**	−7.516**	−4.521**
	(0.058)	(0.320)	(0.188)
top_stars	0.315**	1.598**	0.962**
	(0.010)	(0.054)	(0.031)
lspline(distance, c(1.2, 3))1	−0.040	−0.218	−0.135*
	(0.021)	(0.112)	(0.067)
lspline(distance, c(1.2, 3))2	0.022*	0.101*	0.055*
	(0.009)	(0.045)	(0.027)
lspline(distance, c(1.2, 3))3	−0.243**	−1.272**	−0.742**
	(0.049)	(0.274)	(0.159)
lnprice	0.274**	1.394**	0.839**
	(0.010)	(0.054)	(0.032)
weekend	0.093**	0.533**	0.316**
	(0.010)	(0.051)	(0.030)
Num.Obs.	11 397	11 397	11 397
R2	0.230		
PseudoR2		0.188	0.188

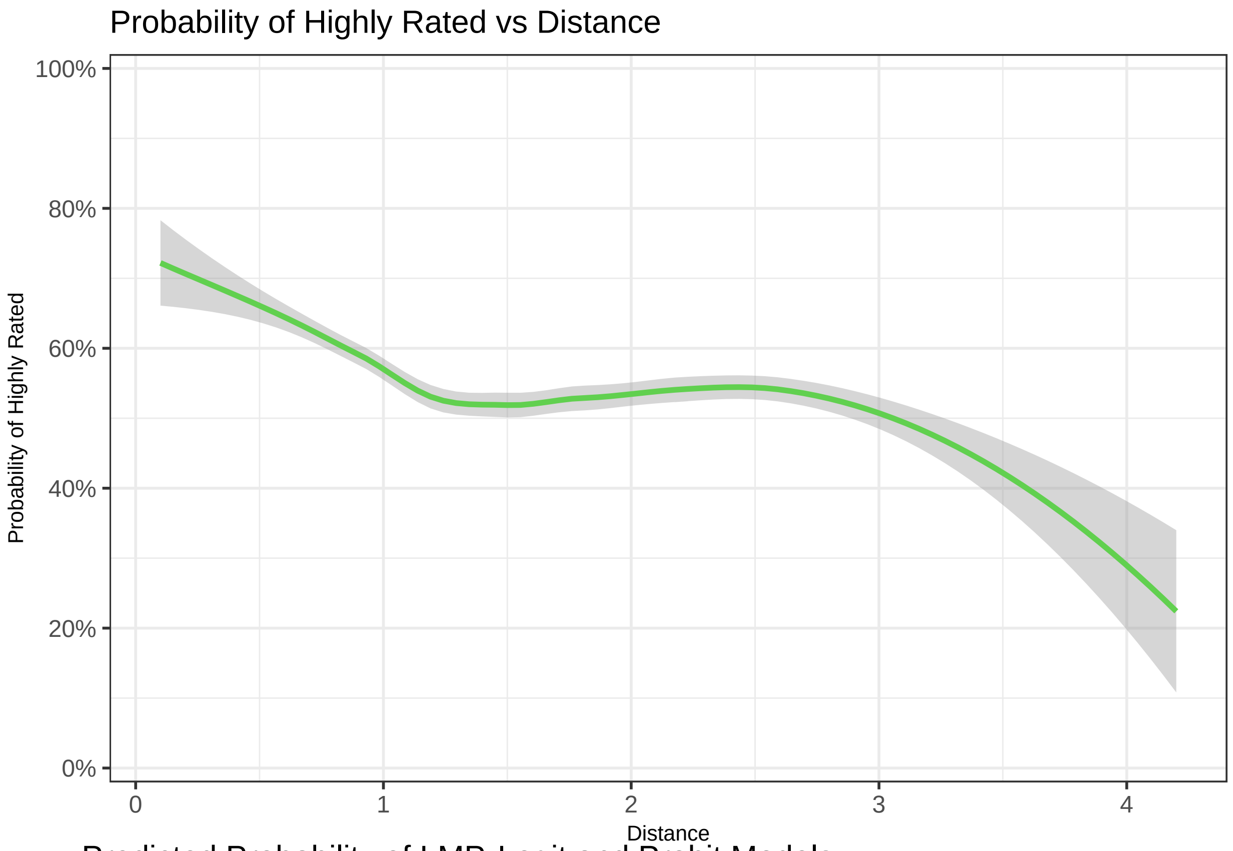
* p < 0.05, ** p < 0.01

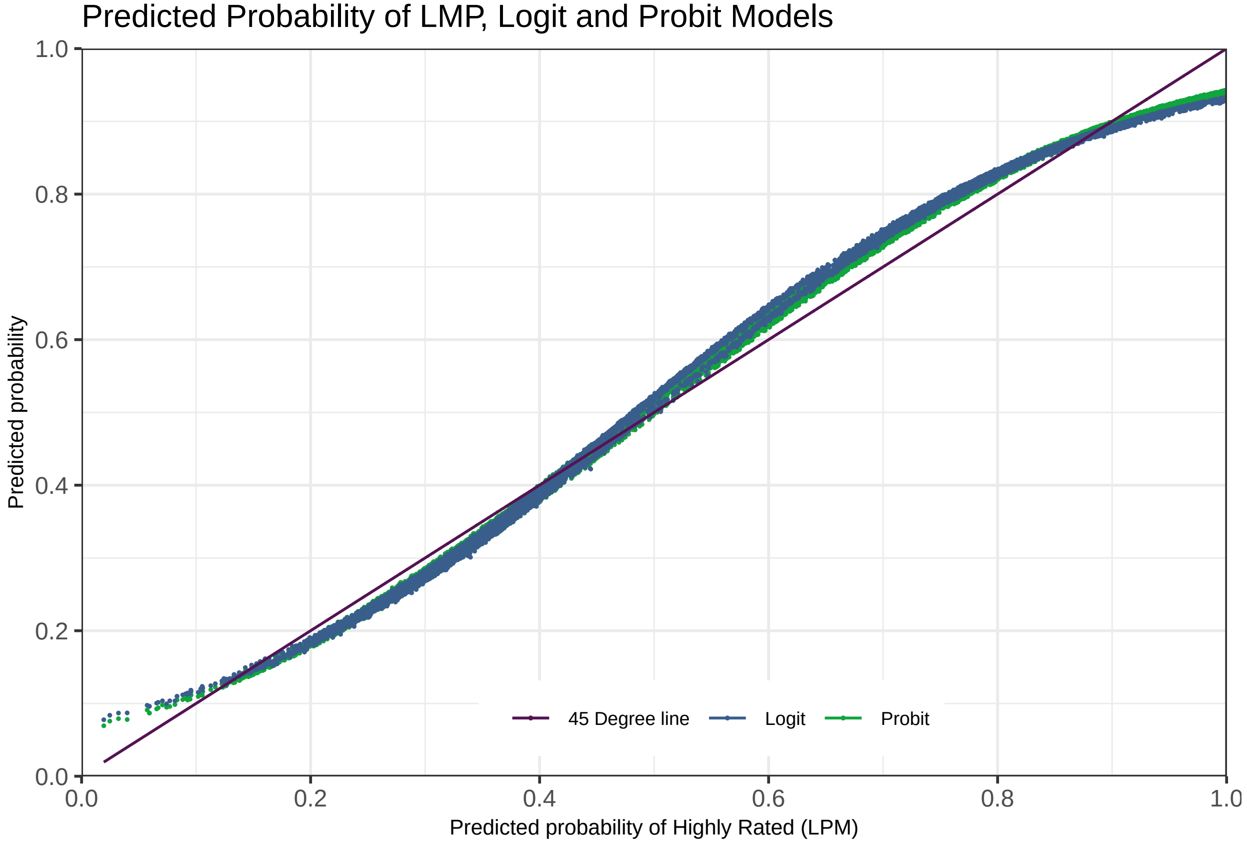
```
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
```

```
#distance
g11 <- ggplot(data = data, aes(x=stars, y=highly_rated)) +
  geom_smooth(method="loess", color="3a5e8cFF") +
  scale_y_continuous(expand = c(0.01,0.01),limits = c(0,1), breaks = seq(0,1,0.2), labels = scales::perc
  labs(x = "Distance",y = "Probability of Highly Rated") +
  theme_bw() +
  ggtitle("Probability of Highly Rated vs Distance") +
  theme(plot.title = element_text(size = 12), axis.title = element_text(size=8) )
g11
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 3

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at 3
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius 1

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 0

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 1

## Warning: Removed 20 rows containing missing values (geom_smooth).

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```




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


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

Appendix

- 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