Group 7: Jialiang Cui, Lara Daoud, Siow Meng Low, Rémi Sabonnadiere

Digital Marketing Analytics

Group Project

Introduction

This assignment uses the data from Expedia Kaggle competition to study hotel listing strategy on Expedia search results (Kaggle, 2017). The main content of this report is within 20 pages, the remaining pages are for bibliography and appendices. The appendices include supplementary materials such as regression results, supplement figures and tables.

Part 1, Question 1: Effects on Conversion Rate by Listing of Hotels

Methodology

When user searches for a hotel in Expedia, usually he or she will get a sorted result of hotels. However, users may get random results in some cases. In this dataset, 277,797 unique searches are sorted and 121,545 searches are random. An intuitive way to test the effectiveness of sorting result is to calculate the average conversion rate. Conversion rate is measured by bookings/clicks. The conversion rates of individual searches are first calculated and they are then aggregated based on if random sorted or not. The result is following:

- Random listing: 55%
- Sorted listing: 94%

Z-test of proportions has been performed and this difference is statistically significant. In conclusion, sorted listing will greatly increase the conversion. If sorting has significant impact on conversion rate, listings order will most likely to be important. To measure how listing order affects user decisions, the data needs to be grouped by position, which indicates the hotel position in search result.

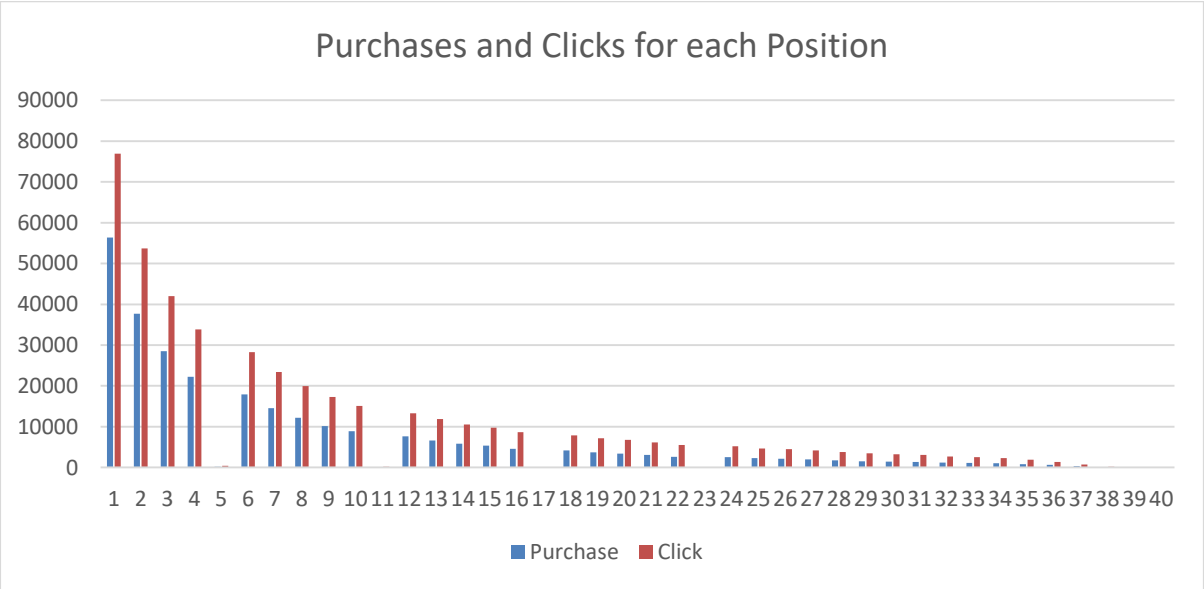

Figure 1 Number of Purchases and Clicks for each Position

As shown in the figure, hotels with higher positions will generate more clicks and more likely to be booked. In general, position has a strong correlation with click or purchase/booking. In order to quantitatively measure

how position will affect the booking, a logistic regression is performed. Logistic regression is suitable to capture the probability of booking given the fact that booking is a binary variable.

To get a more accurate estimate, variables other than position need to be controlled for. There are two types of variables included in the model: static hotel characteristics and dynamic hotel characteristics. Static characteristics include hotel rating, hotel review score, if hotel belongs to a hotel chain, and hotel location score. Dynamic characteristics include position, price and promotion flag. Intuitively these variables are relevant to users' decision making. Since variables such as price and rating can be the metrics of sorting, it is necessary to check if multicollinearity exists in the model. VIF (variance inflation factor) test is used to test the multicollinearity. The result turns out to be no multicollinearity occurs as all variables have low VIF score (less than 3).

Note that in the following logistic regression, price_usd is transformed using logarithm. This is because the hotel price range is wide and log transformation normalises this. Also, log transformation allows us to analyse the effects of price in percentage terms (instead of absolute term) in the regression equation.

Logit Regression Results

Dep. Variable:	booking_bool	No. Observations:	9902847
Model:	Logit	Df Residuals:	9902839
Method:	MLE	Df Model:	7
Date:	Thu, 15 Jun 2017	Pseudo R-squ.:	0.1110
Time:	15:54:09	Log-Likelihood:	-1.1215e+06
converged:	True	LL-Null:	-1.2615e+06
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
const	-1.0479	0.016	-64.698	0.000	-1.080	-1.016
prop_starrating	0.1006	0.002	41.222	0.000	0.096	0.105
prop_review_score	0.1751	0.003	67.555	0.000	0.170	0.180
prop_brand_bool	0.1553	0.004	36.077	0.000	0.147	0.164
position	-0.1155	0.000	-407.690	0.000	-0.116	-0.115
log_price_usd	-0.4825	0.004	-132.138	0.000	-0.490	-0.475
promotion_flag	0.1494	0.005	33.069	0.000	0.141	0.158
prop_location_score1	0.0120	0.001	8.263	0.000	0.009	0.015

Figure 2 Logistic Regression Results to Analyse Listing Position

The regression shows all variables are significant. The estimates of logistic regression coefficients are determined by maximising the log likelihood. The equation of the regression can be expressed using log odds:

$$\log\widehat{(odds)} = \widehat{\log}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

p: the probability that user will book

Odds ratio measures the relative probability, the probability that event will happen against the probability the event won't happen.

$$odds(p) = \frac{p}{1-p}$$

Logistic regression gives the marginal odds for each independent variable.

Variable	Odds	[95% Confidence Interval]	
(Intercept)	0.3507	0.3396	0.3620
prop_starrating	1.1058	1.1008	1.1107
prop_review_score	1.1914	1.1853	1.1972
prop_brand_bool	1.1680	1.1584	1.1782
position	0.8909	0.8905	0.8914
price_usd	0.6172	0.6126	0.6219
promotion_flag	1.1611	1.1514	1.1712
prop_location_score1	1.0121	1.0090	1.0151

Table 1 Marginal Odds of Each Variable

When other variables are controlled for, position has the marginal odds of 0.8909. It can be interpreted as that the odds ratio of booking will decrease by 10.91% if there is one unit increase in position (for instance, hotel used to be the first on the listing now becomes the second). This result is consistent with the findings that higher position brings more bookings.

Part 1, Question 2: Geographically Diverse or Diverse in Price/Quality

Approach

Two approaches were taken to answer this question, with the difference being what is defined as a user response. The first definition of user response is whether or not a user booked, while the second definition is how many links were clicked on by the user. This made it possible to get more insight into how different aspects of user behaviour are affected by the diversity of options.

Assumptions

Every search ID was assumed to be a different user in the data. While this results in the loss of some important information such as users that left a search and came back and made a booking at a later time, it was a necessary assumption since no data on user ID was available.

Methodology

Records were grouped by search ID (under the assumption that every search was a different user). How diverse the results are in terms of a specific criterion was defined as the standard deviation of the corresponding variable. For example, the standard deviation of the price variable was used to measure how diverse in price the results are.

The property star rating was used as a measure of quality. Since no data was provided about the geographical distance between hotels, geographical distance from the user was used to calculate how geographically diverse the search results were by taking the standard deviation of the geographical distance between the user and the properties. The accuracy of this calculation would depend on how far the user is from the destination in which they are searching for properties in. If the user is close to or even already in the destination and the geographical distance between user and property is used as a measure of geographical diversity between properties, the measure would be less accurate and would likely underestimate the geographical diversity of the results than if the destination was further away. For example, suppose a user was already in Paris and is looking for a hotel in Paris. If two search results are shown, one of a hotel 10 km south from the user and another hotel that is 10 km north of the user. Using this method, the geographical diversity between these two hotels is 0, while in reality they are indeed geographically diverse.

Another limitation of this approach is that a user can search for hotels in different cities in one search, or perform a search by hotel name and not by a specific area. This will increase the value of the standard deviation and overestimate the geographic diversity of the results.

Models

Two models were created for each of the definitions of user response; one model had the diversity of the price to quality ratio of as one of the independent variables, the second model had two separate variables to include how diverse the search results were in terms of price and quality. In summary, there are four different models:

- Model 1
    - Response measure: booking boolean
    - Diversity measures: geographically diverse, price/quality ratio diverse
- Model 2
    - Response measure: booking boolean
    - Diversity measures: geographically diverse, price diverse, quality diverse
- Model 3
    - Response measure: number of clicks
    - Diversity measures: geographically diverse, price/quality ratio diverse
- Model 4
    - Response measure: number of clicks
    - Diversity measures: geographically diverse, price diverse, quality diverse

Results

The results of models 1, 2, 3 and 4 are shown in Table 15, Table 16, Table 17, and Table 18 respectively in the Appendix. In order to gain more insight into whether the size of an estimated coefficient is economically significant or not, it is important to look into the typical value of the variable being examined. Variables might seem economically insignificant if the estimated coefficient is small, but become more significant if the typical value of that variable is very high. The expected values of each of the variables of interest are shown in the table below.

Variable	Expected Value
Search results' geographic diversity	3.35
Search results' price diversity	131.06
Search results' quality diversity	0.78
Search results' price/quality ratio diversity	17.68

Table 2 Expected Value of Variable of Interests

Geographically diverse results: the results show that whether a booking boolean or number of clicks is used as the definition of user response, the geographic diversity of the results is statistically significant with a negative effect on response rate. When the response variable is defined as whether or not a booking is made, the odds of booking is about 4-6% less if geographic diversity increases by 1 unit. The effect is much smaller if number of clicks is taken as the definition of user response. This can be due to several reasons, one example is that users, who are browsing where to stay in the area, are not affected as much by results that are geographically diverse. Another reason could be that factors other than geographic location are more important for users when deciding whether or not to purchase a hotel room.

Figure 3 shows the number of bookings resulting from search results of different levels of geographic diversity.

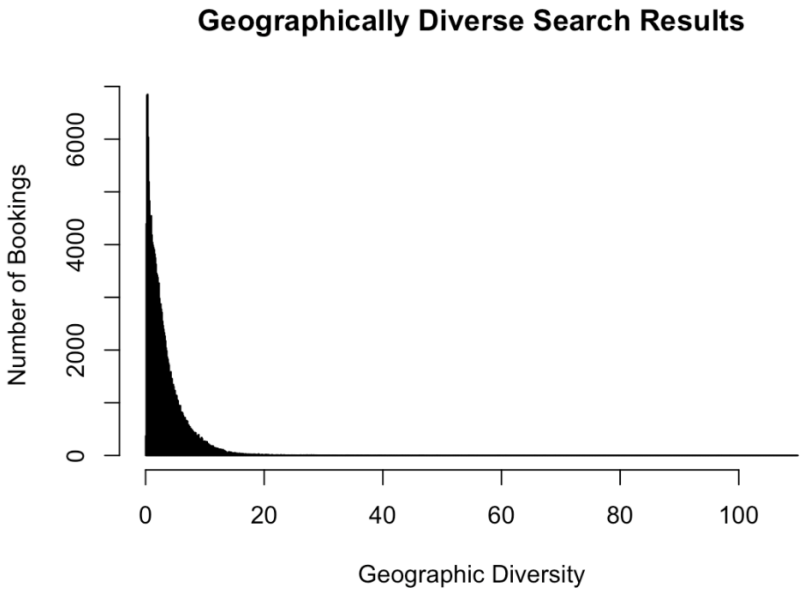

Figure 3 Number of Bookings against Geographic Diversity

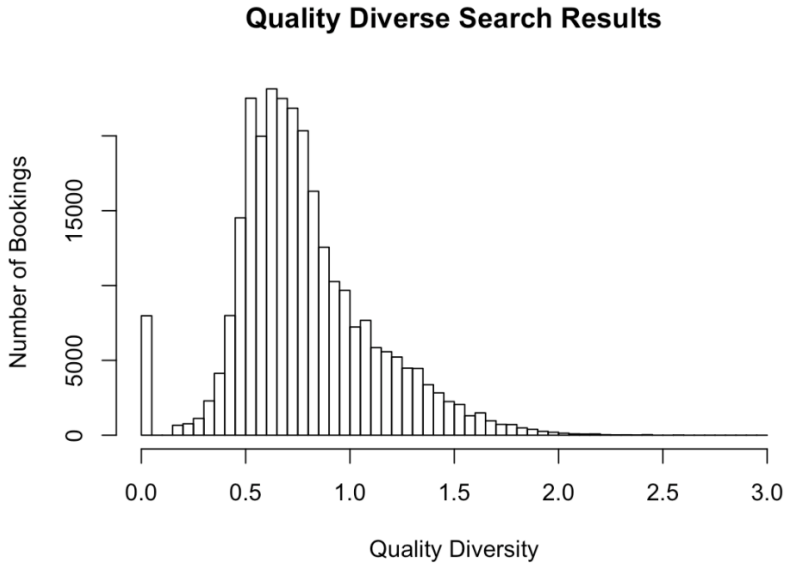

Figure 4 Number of Bookings against Quality Diversity

Quality diverse results: results show that search results that are diverse in terms of quality of hotels shown are significant both when the dependent variable is a booking boolean or number of clicks. One unit increase in diversity of quality increases the odds of booking by around 5%, but decreases the odds of clicking on the results. This could be due to the fact that users are more confident about making a booking when they see how the quality of the hotel they have chosen compares to others hotels. The implication of this result to the company depends on what the goal of the company is. If the main objective of the company is to increase bookings made, it is recommended that it shows results with hotels of different star ratings. However, it is possible that the company might want to encourage users to look at more hotels in more detail by clicking on their links and browsing what amenities are available. In this case, the recommended course of action is to show hotels that are of similar star ratings to encourage users to look into hotels in more detail.

Figure 4 shows the number of bookings done resulting from search results of different levels of quality diversity.

Price diverse results: the results show that how diverse the search results are in terms of price is statistically significant, but has a very low estimated coefficient and might therefore not be economically significant only when booking boolean is used as the definition of user response, with a negative effect on the likelihood that a booking is made. However, the expected value of the level of price diversity is relatively higher than the other variables, meaning that price diversity can also be economically significant if it is high enough. The low estimated coefficient could be due to the fact that users make their search with a specific amount they are willing to spend in mind, and hence their willingness to make a booking and their interest in browsing hotels (some with prices that are not within the amount they are looking to spend) are not affected by seeing hotels that are of different prices ranges.

Price/quality ratio diverse results: the results show that how diverse the search results are, in terms of the price to quality ratio, is statistically significant only when the response variable is defined as whether or not a booking was made, where search results that are diverse in terms of price/quality ratio decrease the likelihood that a booking is made. Again, the low estimated coefficient makes price/quality ratio less economically significant, especially since the variable has low typical expected values.

Part 1, Question 5: Tailoring Results Based On Characteristics of User Session

As seen in earlier section, different property related variables have different effects on the probability of booking. For instance, higher star rating / review score / location score have positive effects on the probability of booking. Sale promotions or being a hotel chain also helps to improve the probability of booking. As expected, a higher hotel price will also lead to lower chance of booking (controlling for other factors). In this section, we will look at some other user related attributes (specifically search attributes) and investigate how they affect the probability of booking.

Methodology

Logistic regression will again be used to estimate the effects of independent variables. Three user search attributes will be added to the logistic regression, namely the length of stay, room counts, and children counts specified by user in the search. Note that a large percentage of users do not have historical spending information, hence these variables are not added into the regression.

Some interaction effects with these user search attributes have also been added to the regression, the results are shown in Figure 5. One thing to note is that after controlling for the interaction effects between length of stay and location score, the effects of location score alone became statistically insignificant. Other than that, the coefficient signs of the hotel variables largely remain the same. For example, if promotion flag is true, there will be around 10% increase in the odds of booking. If a hotel belongs to a big hotel chain, there will be about 13% increase in the odds of booking. If the hotel review score increases by one point, odds of booking will increase by 18%. In short, these hotel characteristics have significant impacts on users' decision making.

Next, we will examine the effects of the user search characteristics and their interactions with other variables:

Number of children specified in the search: On average, increasing number of children by 1 in the search decreases the odds of booking by about 15% (marginal odds = $e^{-0.1590} = 0.853$). The interesting part is its effects with other price, star rating, and brand_bool. The coefficients of these interaction terms are all positive, implying that users who travel with children would not mind paying for higher hotel price (since the gradient of *log(price)* becomes less negative with increasing number of children, controlling for other factors. In addition, travellers with children seem to prefer hotels with higher star ratings and hotel chains, controlling for other factors. All these are intuitively explainable, travellers with children would prefer hotel with amenities

and thus higher star hotel (or reputable hotel chains) are preferred. This group of travellers also tend to be less sensitive to price, compared to traveller without children. Out of these three interaction terms, the interaction between number of children and star rating seems to be the strongest. Therefore, Expedia could recommend hotel with higher star ratings to travellers who travel with many children.

Number of rooms specified in the search: The coefficient of "srch_room_count" is statistically insignificant. This means that controlling for all other factors, the number of rooms required does not affect the probability of booking. The interaction between room count and *log(price)* is also statistically insignificant. Hence we say that the room count specified by users does not seem to affect probability of booking.

Logit Regression Results

Dep. Variable:	booking_bool	No. Observations:	9902847
Model:	Logit	Df Residuals:	9902829
Method:	MLE	Df Model:	17
Date:	Thu, 15 Jun 2017	Pseudo R-squ.:	0.1135
Time:	18:58:37	Log-Likelihood:	-1.1183e+06
converged:	True	LL-Null:	-1.2615e+06
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
const	-0.8049	0.044	-18.314	0.000	-0.891	-0.719
prop_starrating	0.1008	0.003	37.280	0.000	0.095	0.106
prop_review_score	0.1663	0.003	63.922	0.000	0.161	0.171
prop_brand_bool	0.1213	0.005	25.548	0.000	0.112	0.131
position	-0.1152	0.000	-406.247	0.000	-0.116	-0.115
log_price_usd	-0.4801	0.009	-51.258	0.000	-0.499	-0.462
promotion_flag	0.0996	0.007	13.537	0.000	0.085	0.114
prop_location_score1	-0.0026	0.002	-1.117	0.264	-0.007	0.002
srch_length_of_stay	-0.1193	0.008	-15.198	0.000	-0.135	-0.104
srch_children_count	-0.1590	0.022	-7.263	0.000	-0.202	-0.116
srch_room_count	-0.0158	0.032	-0.486	0.627	-0.079	0.048
child:lprice	0.0211	0.005	4.231	0.000	0.011	0.031
child:star	0.0216	0.003	6.758	0.000	0.015	0.028
child:brand	0.0200	0.006	3.435	0.001	0.009	0.031
room:lprice	0.0114	0.007	1.698	0.089	-0.002	0.025
los:lprice	-0.0027	0.002	-1.664	0.096	-0.006	0.000
los:location	0.0101	0.001	12.097	0.000	0.008	0.012
los:promo	0.0367	0.003	14.559	0.000	0.032	0.042

Figure 5 Logistic Regression Results to User Search Characteristics

Length of stay specified in the search: Each additional nights specified in the length of stay by user decreases the odds of booking by about 11% (marginal odds = $e^{-0.1193} = 0.888$). Its interaction with *log(price)* is statistically insignificant, therefore there is no evidence showing that traveller, who requires longer stay, is more or less sensitive to hotel price. However, its interactions with location score and promotion_flag have positive coefficients. This means that a traveller who stays for higher number of nights would prefer to stay in hotels with good location score, controlling for other factors. In addition, this traveller would also prefer a hotel that has sale price promotions, keeping all other independent variables constant. Out of these two interactions, the interaction with promotion_flag is stronger. Therefore a possible strategy to tailor search results to travellers who require longer stay, is to prioritise hotels that have sale price promotions (and preferably have good location scores).

Part 1, Question 6: Tailoring Results Based On Expedia Sites

The original question asks how the results should be tailored based on the channel (e.g. search engine, tripadvisor.com) customer arrived to Expedia. Since the Kaggle dataset does not have the information on the channels, we use "Site ID" as a proxy to answer this question.

Site ID refers to the ID of the Expedia point of sale (e.g. Expedia.com, Expedia.co.jp). This section investigates whether there are different preferences of users arriving to these sites and recommends any possible tailoring.

Identifying the Sites Needing Improvement

There are 34 distinct site IDs in the dataset, in order to identify which site most urgently needs improvement via tailoring, we will need the booking rate (number of bookings per impression) and the click-through-rate (number of clicks / number of impressions), break down by site ID. Table 3 lists down the 8 sites with the lowest booking rate.

Note that in order for the booking rate and CTR to be representative, sites with less than 10000 search results are filtered away from this table.

Site ID	Booking per Impression	Click-Through-Rate
19	0.0207	0.0581
10	0.0218	0.0459
9	0.0235	0.0480
25	0.0236	0.0477
4	0.0236	0.0487
17	0.0239	0.0564
13	0.0244	0.0528
16	0.0248	0.0470

Table 3 The 8 Sites with Lowest Booking Rate

Site 13, 17 and 19 have relatively high CTR but low booking per impression, hence they are the ones that require improvement in the booking rate. The following sections focus on these three sites and examine the possible tailoring methods.

Demographics of the Visitors

In the appendix, Figure 17 shows the number of unique countries of the Expedia site visitors and Figure 18 depicts the number of unique destinations searched by visitors to each site. It can be seen that the visitors to site 13, 17 and 19 primarily came from a small set of countries and the number of unique destinations they searched are also relatively few.

This implies that Expedia can potentially tailor the results shown on this site to suit the demographics of their visitors. For instance, say the site is primarily used by Japanese visitors and the search destination is mainly Western Europe, Expedia can potentially tailor the search results such that the hotel with staff who speaks Japanese would rank higher in the search results. This could potentially improve the probability of booking.

Receptiveness to Promotion and Hotel Chain

If the visitors to these 3 sites respond better to sale price promotions or hotel chains (instead of independent hotels), the search results can potentially be tailored to prioritise them. To affirm this, logistic regression will be used to examine whether the visitors to these 3 sites have such preferences.

The target variable is booking_bool (binary variable), and we add the 3 dummies (site 13, 17 or 19) as independent variables and also their interaction terms with promotions and brand_bool. Other variables (e.g. review score, position, price in log format) are also added as control variables. The result of the regression is shown in Figure 6. The results show that the three interaction terms with promotions are all statistically insignificant. This means the visitors to these three sites do not have statistically different response rate to sale promotions, compared with the other users. Hence promotions cannot be used for tailoring the search results of these 3 sites.

For the interaction terms with brand_bool, all three coefficients are statistically significant. The interaction between site13 and brand_bool is negative while the other two coefficients are positive. This implies that on average site 13 visitors prefers independent hotels while site 17 & 19 visitors prefer hotel chains. Expedia can tailor search results accordingly (i.e. show more independent hotels on site 13 while display more hotel chains in site 17 and 19).

Logit Regression Results

Dep. Variable:	booking_bool	No. Observations:	9902847
Model:	Logit	Df Residuals:	9902829
Method:	MLE	Df Model:	17
Date:	Thu, 15 Jun 2017	Pseudo R-squ.:	0.1113
Time:	13:16:48	Log-Likelihood:	-1.1211e+06
converged:	True	LL-Null:	-1.2615e+06
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
const	-0.9828	0.017	-59.437	0.000	-1.015	-0.950
site13	-0.2036	0.073	-2.803	0.005	-0.346	-0.061
site17	-0.5316	0.098	-5.427	0.000	-0.724	-0.340
site19	-0.6535	0.037	-17.430	0.000	-0.727	-0.580
prop_starrating	0.1075	0.002	43.637	0.000	0.103	0.112
prop_review_score	0.1713	0.003	65.789	0.000	0.166	0.176
prop_brand_bool	0.1482	0.004	34.264	0.000	0.140	0.157
prop_location_score1	0.0127	0.001	8.728	0.000	0.010	0.016
prop_log_historical_price	-0.0048	0.001	-4.379	0.000	-0.007	-0.003
position	-0.1156	0.000	-407.769	0.000	-0.116	-0.115
log_price_usd	-0.4916	0.004	-133.107	0.000	-0.499	-0.484
promotion_flag	0.1482	0.005	32.631	0.000	0.139	0.157
site13:promo	-0.0349	0.102	-0.343	0.732	-0.234	0.165
site17:promo	0.0713	0.121	0.590	0.555	-0.166	0.308
site19:promo	0.0127	0.051	0.248	0.804	-0.088	0.113
site13:brand	-0.2181	0.110	-1.989	0.047	-0.433	-0.003
site17:brand	0.2519	0.125	2.012	0.044	0.007	0.497
site19:brand	0.2264	0.051	4.435	0.000	0.126	0.326

Figure 6 Logistic Regression Results with Site ID Dummies

Extra: Tailor Site Homepage to Show Last-Minute Deals

The variable "srch_booking_window" provides information on the number of days between the search date and the hotel stay date. It is discovered that the average number of days between search and travel for site 19 is significantly lower than the global average (T-test results are shown in Table 4).

Overall Average Number of Days Between Search Date and Travel Date	37.62
Average Number of Days Between Search Date and Travel Date (Site 19)	25.82
Two-sample T-Test of Means (Two-Sided)	T-Statistics: 93.23 P-Value: 0.0

Table 4 Average Number of Days between Search Date and Travel Date

The above implies that on average the visitors to site 19 are planning to travel sooner than the other sites. Expedia can potentially tailor site 19 homepage such that more last-minute deals are shown, in order to increase the probability of customer booking.

Part 2a: Identifying Competitors for Hotel Chain

The dataset is related to Expedia searches. To a hotel chain, the "competitors" in this context would be the other hotels which show up in the same search. In the "Descriptive Statistics" section, we will investigate the characteristics of the competing hotels.

Descriptive Statistics

In this section, we will explore the characteristics of the competing hotels and their distribution in order to know how the competing hotels look like. In the Expedia dataset, there are 53,115 unique hotel chains. A thousand samples will be randomly drawn from these 53,115 hotel chains and their competitors identified through the relevant searches.

The Expedia search results which include these 1000 sample hotel chains are extracted and the competitors (which show up in the same search results) are then analysed. The competitors' characteristics are described in the subsequent sections.

Competitors' Country Location

In these sample searches, 99.55% of them contain only hotels from a single country. The other 0.45% contains hotels from two distinct countries. In other words, the competitors are predominantly located in the same country as the hotel chain.

Hotel Star Rating Difference between Hotel Chains and Competitors

Each hotel chain has different star rating and we are interested in the amount of similarity between a hotel chain and its competitors. Therefore, we subtract a hotel chain's star rating from its competitors' star ratings. Table 5 shows the statistics of this difference and Figure 7 depicts the distribution of this difference.

Median	0
Mean	-0.062
Standard Deviation	1.031
Percentage of Hotels with same Star Rating	45.39%
Percentage of Hotels with Star Rating Difference ≤ 1	88.75%

Table 5 Descriptive Statistics of Hotel Star Rating Difference between a hotel chain and its competitors

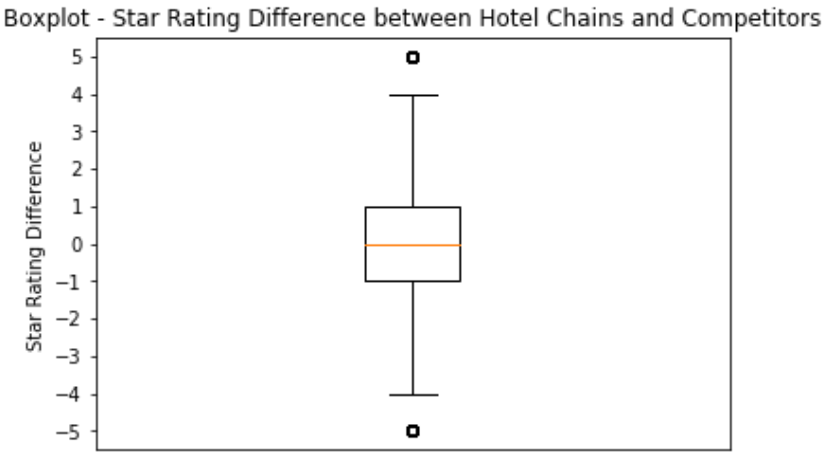

Figure 7 Boxplot Depicting Distribution of Star Rating Difference

As observed, most of the competitors (almost 90%) are no more than one star rating difference from the hotel chains. Therefore, the primary competitors of a hotel chain are those with no more than one star difference. For instance, the primary competitors of a 3-star hotel would be 2-star, 3-star and 4-star hotels.

Review Score Difference between Hotel Chains and Competitors

Similar to hotel rating, we are interested in the review score similarity between a hotel chain and its competitors. Thus, we subtract the hotel chain's review score from its competitor's scores to analyse the difference. The results are shown in Table 6 and Figure 8.

Note that some of the hotels do not have review score yet. These hotels are excluded from this analysis.

Median	0
Mean	-0.055
Standard Deviation	1.062
Percentage of Hotels with same Review Score	31.94%
Percentage of Hotels with Review Score Difference ≤ 0.5	70.10%
Percentage of Hotels with Review Score Difference ≤ 1	86.44%

Table 6 Descriptive Statistics of Hotel Review Score Difference between a hotel chain and its competitors

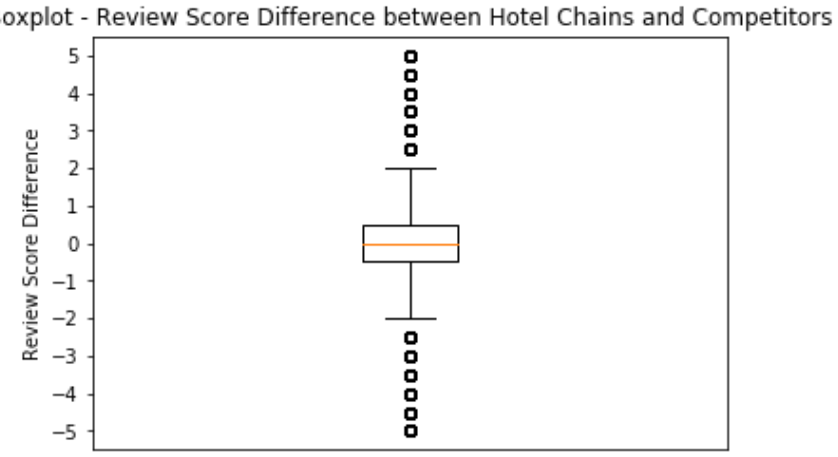

Figure 8 Boxplot Depicting Distribution of Review Score Difference

More than 70% of the competitors have review score difference less than or equal to 0.5. A hotel chain's competitors are primarily those with no more than 0.5 review score difference. For example, the primary competitors of a hotel chain with 4.5 review scores would be the hotels with review scores of 4, 4.5 or 5.

Hotel Types of Competitors

Next, we examine the hotel types of the competitors. To perform this, we look at the proportions of hotel chains (brand_bool = 1) in the sample searches and compare it against the proportions observed in the overall data. Table 7 tabulates the proportions observed in these two scenarios.

Number of Hotels in all Searches	9,917,530
Number of Hotel Chains in all Searches	6,290,731
Proportion of Hotel Chains in all Searches	6,290,731 / 9,917,530 = 63.43%
Number of Hotels in Search Results Pertaining to 1000 Sample Hotel Chains	3,537,460
Number of Hotel Chains in Search Results Pertaining to 1000 Sample Hotel Chains	2,715,925
Proportion of Hotel Chains in Searches Pertaining to	2,715,925 / 3,537,460 = 76.78%

1000 Sample Hotel Chains	
Two-sided Z-test of Proportions	Z-Statistics = -458.09 P-Value ≅ 0.0

Table 7 Proportions of Hotels that are Hotel Chains in Overall Searches and in Sample Searches

It can be seen that the percentage of hotel chains in the sample searches (search results containing the 1000 sample hotel chains) is 76.78% and it is quite different from the percentage in all searches (63.43%). A two-sided Z-test of proportions proves that these two proportions are statistically different.

In summary, as a hotel chain, my competitor is more likely to be another hotel chain than an independent hotel.

Location Score Difference between Hotel Chains and Competitors

To analyse the location score, we again subtract a hotel chain's location score ("prop_location_score1" in the dataset) from its competitors. The distribution of this difference is detailed in Table 8 and Figure 9.

As we can observe from Figure 9, there are quite a number of outliers with huge location score differences. However, most of the competitors are still concentrated in the region where the location score difference is no more than 1.5.

Thus, we conclude that a hotel chain's primary competitors are those with location score difference less than or equal to 1.5. For instance, the competitors of a hotel chain with location score 3.0, would be those with location score between 1.5 and 4.5.

Median	0.0
Mean	0.063
Standard Deviation	1.204
Percentage of Hotels with same Location Score	11.43%
Percentage of Hotels with Location Score Difference ≤ 0.5	50.25%
Percentage of Hotels with Location Score Difference ≤ 1	69.35%
Percentage of Hotels with Location Score Difference ≤ 1.5	80.97%
Percentage of Hotels with Location Score Difference ≤ 2	89.47%

Table 8 Descriptive Statistics of Hotel Location Score Difference between a hotel chain and its competitors

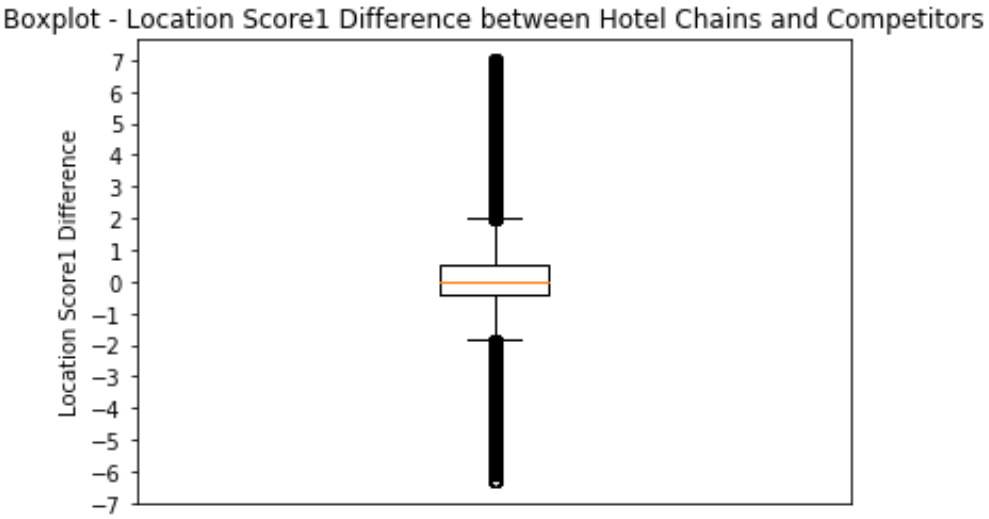

Figure 9 Boxplot Depicting Distribution of Location Score Difference

Price in USD (Ratio)

This subsection investigates the relationship between a hotel chain's price and its competitors'. Note that hotels of different price range (e.g. budget, mid-range, luxurious) may price their rooms vastly differently. It is therefore more effective to use the price ratio (i.e. competitor's price divided by hotel's price). Using the price ratio, we can examine the price difference in terms of percentages (e.g. number of competitors who have 25% price difference compared to my price) instead of absolute dollar values.

Median Price Ratio	1.0
Mean Price Ratio	1.143
Standard Deviation	1.369
Percentage of Competitors with Price Difference ≤ 25%	44.14%
Percentage of Competitors with Price Difference ≤ 50%	72.20%
Percentage of Competitors with Price Difference ≤ 75%	86.54%

Table 9 Descriptive Statistics of Price Ratio (Competitor's Price / Hotel's Price)

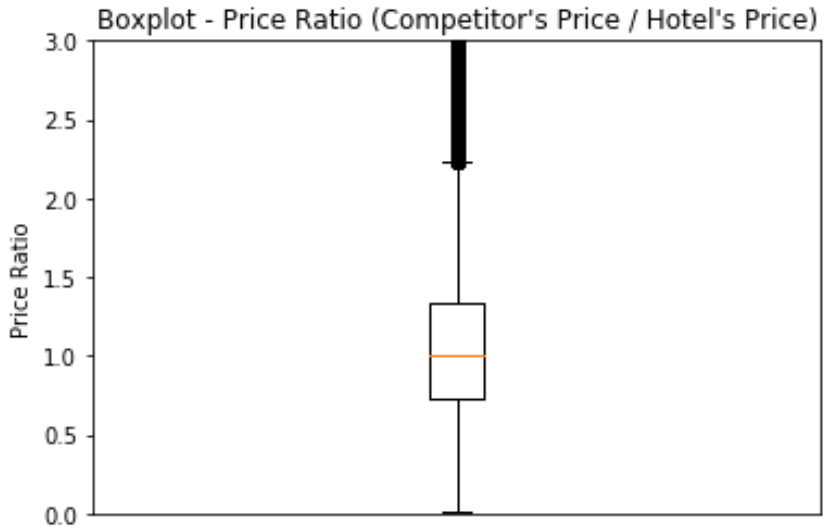

Figure 10 Boxplot Depicting Distribution of Price Ratio

Note that in Figure 10, the outliers with Price Ratio > 3.0 are not shown. From the boxplot, it can be seen that vast majority of the competitors price their rooms within 50% difference. Therefore, we can say that a hotel chain's primary competitors are those with price difference no greater than 50%.

Historical Price Ratio

This subsection investigates the relationship between a hotel chain's historical price and its competitors'. Note that the Expedia dataset provides historical price in log format hence we need to exponent it.

Similar to earlier reasoning, we use historical price ratio (competitor's historical price / hotel chain's historical price) to examine the price difference in terms of percentages (e.g. number of competitors who have 25% historical price difference compared to mine).

Median Price Ratio	1.0
Mean Price Ratio	1.205
Standard Deviation	6.062
Percentage of Competitors with Historical Price Difference ≤ 25%	53.29%
Percentage of Competitors with Historical Price Difference ≤ 50%	79.19%
Percentage of Competitors with Historical Price Difference ≤ 75%	90.39%

Table 10 Descriptive Statistics of Historical Price Ratio (Competitor's Historical Price / Hotel's Historical Price)

From Figure 11, more than three quarters of my competitors are having historical price difference less than 50%. Hence we conclude that our primary competitors are those with historical price difference less than 50% (when compared to our historical price).

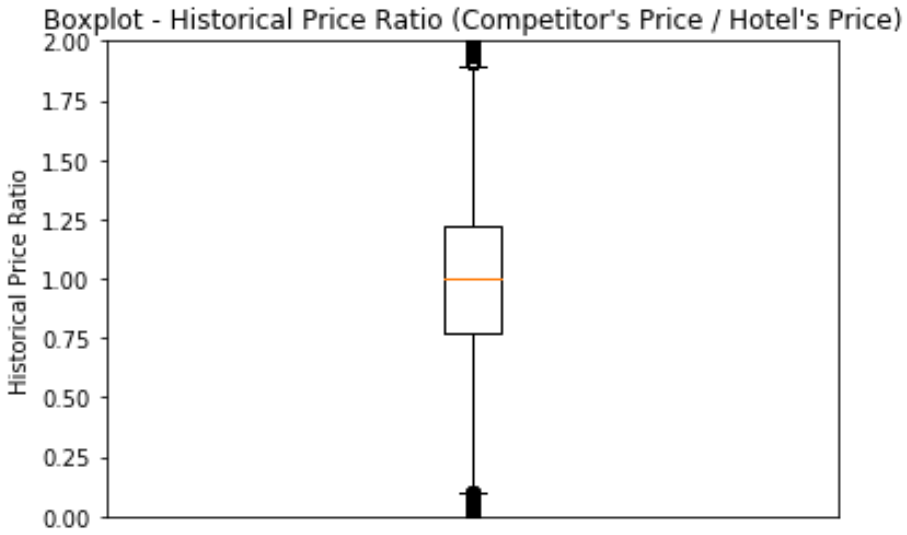

Figure 11 Boxplot Depicting Distribution of Historical Price Ratio

Competitors with Promotions

In this section, we will investigate whether a hotel chain's competitors are likely to have sale price promotions on Expedia. For this purpose, we analyse two scenarios:

1. Scenario 1: My hotel chain is having sale price promotions, what percentage of my competitors is also having promotions?
2. Scenario 2: My hotel chain is not having promotions, what percentage of my competitors is having sale price promotions?

In the case of scenario 1, the percentages of promotions are tabulated in Table 11. The percentage of competitors having promotions is 37.95%, and the two-sided Z-test of proportions show that this is significantly higher than the overall percentages of promotions in the whole dataset.

Number of Hotels in all Searches	9,917,530
Number of Promotions in all Searches	2,139,822
Proportion of Hotel Chains in all Searches	2,139,822 / 9,917,530 = 21.58%
Number of Hotels in the Search Results of Scenario 1	555,278
Number of Promotions in the Search Results of Scenario 1	210,718
Proportion of Promotions in Scenario 1	210,718 / 555,278 = 37.95%
Two-sided Z-test of Proportions	Z-Statistics = -284.56 P-Value ≅ 0.0

Table 11 Proportions of Hotels that are Having Promotions in Overall Searches and in Scenario 1

For scenario 2, the results are tabulated in Table 12 and the percentage of competitors having promotions is 17.30%. This proportion is significantly lower than the overall percentages of promotions.

Number of Hotels in all Searches	9,917,530
Number of Promotions in all Searches	2,139,822
Proportion of Hotel Chains in all Searches	2,139,822 / 9,917,530 = 21.58%
Number of Hotels in the Search Results of Scenario 2	2,982,182
Number of Promotions in the Search Results of Scenario 2	515,941

Proportion of Promotions in Scenario 2	515,941 / 2,982,182 = 17.30%
Two-sided Z-test of Proportions	Z-Statistics = 160.10 P-Value ≅ 0.0

Table 12 Proportions of Hotels that are Having Promotions in Overall Searches and in Scenario 2

To summarise, as a hotel chain, if I am having sale price promotions on Expedia, a higher percentage (on average 37.95%) of my competitors (whose hotels are shown in the same search results) will also be having sale price promotions. On the other hand, if I am not having sale price promotions, then my property is less likely to be displayed alongside with properties having sale price promotions.

Summary – Characteristics of Competitors

In summary, the competitors of a hotel chain have the following characteristics:

- Located in the same country
- No more than 1 star rating difference, compared with the hotel chain
- No more than 0.5 review score difference
- High likelihood to be another hotel chain
- No more than 1.5 location score difference
- No more than 50% difference in price
- No more than 50% difference in historical price
- More likely to have sale price promotions if we are having one

Extension to Clustering for Competitor Identification

From the descriptive statistics section, we have discovered that competitors, who show up in the same search results, tend to have similar values for a number of hotel attributes. We can then utilise unsupervised learning approach (e.g. hierarchical clustering) to automatically cluster hotels with similar attributes.

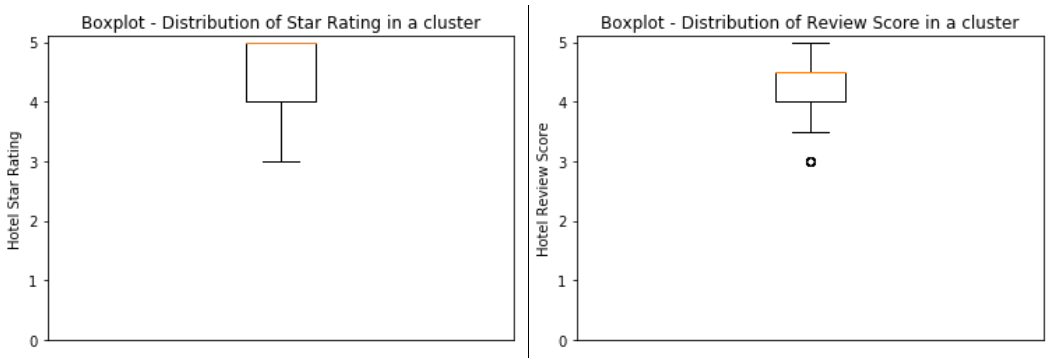

Figure 12 Distribution of Hotel Star Rating and Review Score in a sample cluster

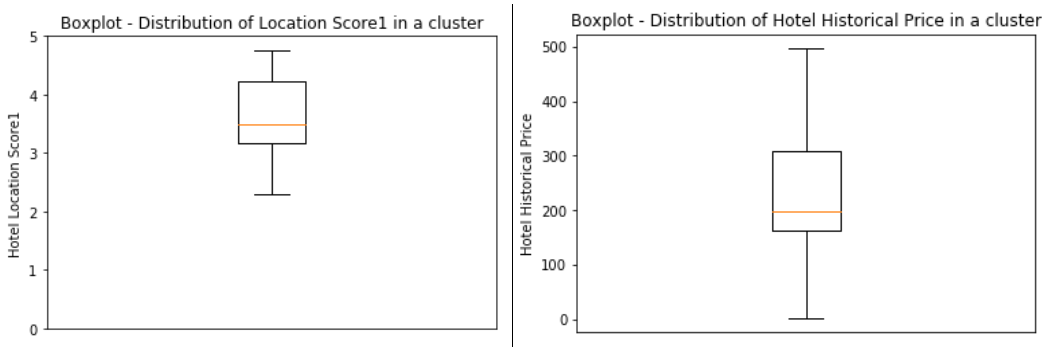

Figure 13 Distribution Hotel Location Score and Historical Price in a sample cluster

Hierarchical clustering (with average linkage) has been performed on hotels with country ID 35. The features to be used for clustering include star rating, review score, brand bool, location score1, historical price, current hotel price and promotion flag. Figure 12 and Figure 13 depict the distribution of attribute values for one of the clusters identified. It can be seen that the hotel within the same clusters (and hence competitors) have very similar attribute value, this is aligned with what is observed in descriptive statistics section.

Hotels with High Conversions per Impression and Click-Through-Rate

Using the Expedia dataset, "booking_bool" and "click_bool" can be used to calculate the "Conversion per Impression" and "Click-Through-Rate" respectively. In addition to have similar characteristics, our strong competitors must also have high historical "Conversion per Impression" and Click-Through-Rate. The table below tabulates the top 5 hotels with the highest Conversions per Impression (note that we only consider hotels with more sufficient samples – those which show up in more than 100 search results).

Property ID	Conversion per Impression	Click-Through-Rate
88359	0.4752	0.5149
66456	0.3484	0.3801
86247	0.3438	0.3906
60275	0.3433	0.3731
29792	0.3306	0.3719

Table 13 Top Five Hotels with High Conversion per Impression (with at least 100 appearances in the dataset)

In summary, clustering approach can be used to separate hotels into different clusters, using the characteristics we identified in descriptive statistics section. As a hotel chain, the other hotels which are in the same cluster of mine would be my competitors. We can then check which hotel, within the same cluster, has high conversion per impression and/or click-through rate. These will be our strongest competitors.

Part 2b: Maximising Probability of Sale

As a hotel chain, I would have to compete against the hotels which show up in the same search and maximise my probability of sale. In the Expedia data, sale is indicated using "booking_bool" (which is equal to 1 if a hotel is booked, 0 otherwise). Since this is a binary variable, logistic regression will be used to estimate the effects of other independent variables on the probability of sale.

Target Variable & Independent Variables

As discussed earlier, the target variable of the logistic regression is:

$$Probability\ of\ Sale = P(booking\_bool\ = 1)$$

To explore how pricing and the characteristics of the hotel listing affect the probability of sale, a number of independent variables are used in the regression.

- (Price Ratio) & $(\text{PriceRatio})^2$
- Brand_bool : Price Ratio (interaction between brand_bool and price ratio)
- Property Star Rating (centred by subtracting the average star rating of competitors)
- Property Review Score (centred by subtracting the average review score of competitors)
- Property brand_bool (indicate whether it is a hotel chain or independent hotel)
- Property Location Score (centred by subtracting the average location score of competitors)
- (Historical Price Ratio) & $(\text{Historical Price Ratio})^2$
- Search results position
- Promotion_flag (indicate if there is any sale price promotions)

The reasoning of selecting these variables as independent variables is described in Table 19 of the Appendix. Note that price ratio is the variable that we are interested at; all other independent variables are added to the regression as control variables.

Other Variables Considered

The following three variables were also considered since they might be correlated with the target variable.

- visitor_hist_starrating: The mean star rating of hotels that a customer has previously purchased may indicate the preferred star rating of hotels that he/she is likely to purchase in future.
- visitor_hist_adr_usd: The mean price per night that the customer has previously purchased may indicate the preferred price range of hotels that he/she is likely to purchase in future.
- srch_query_affinity_score: This represents the log probability of hotel that will be clicked on, based on Expedia's algorithm. A higher probability of being clicked on may also imply higher probability of sale.

However, around 95% of the user searches do not have valid values for these variables. It is also not feasible to impute 95% of the records with a default value and hence, they are not used as independent variables.

Logistic Regression Results

Using the dependent and independent variables discussed in the earlier sections, logistic regression is run and the results are shown in Figure 14.

Logit Regression Results

Dep. Variable:	booking_bool	No. Observations:	9902900
Model:	Logit	Df Residuals:	9902888
Method:	MLE	Df Model:	11
Date:	Thu, 15 Jun 2017	Pseudo R-squ.:	0.1200
Time:	11:34:23	Log-Likelihood:	-1.1101e+06
converged:	True	LL-Null:	-1.2615e+06
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
const	-1.8033	0.018	-100.707	0.000	-1.838	-1.768
price_ratio	-1.4758	0.013	-113.709	0.000	-1.501	-1.450
brand:priceratio	-0.1940	0.013	-15.394	0.000	-0.219	-0.169
price_ratio_sq	0.0577	0.002	29.944	0.000	0.054	0.061
prop_starrating_centred	0.2981	0.003	86.323	0.000	0.291	0.305
prop_review_score_centred	0.1686	0.003	62.491	0.000	0.163	0.174
prop_brand_bool	0.3927	0.012	32.703	0.000	0.369	0.416
prop_location_score1_centred	0.1017	0.002	41.552	0.000	0.097	0.106
historical_price_ratio	0.8330	0.027	30.800	0.000	0.780	0.886
historical_price_ratio_sq	-0.1517	0.010	-14.832	0.000	-0.172	-0.132
position	-0.1100	0.000	-384.714	0.000	-0.111	-0.109
promotion_flag	0.1435	0.004	32.799	0.000	0.135	0.152

Figure 14 Results of Logistic Regression

Effects of Price Ratio

With the above regression results, we can analyse how price ratio affects the probability of sale. For logistic regression model, the estimated odds of sale is modelled as the exponent of the linear function, *f(all)*:

$$Odds\ of\ Sale = \frac{Probability\ of\ Sale}{1 - Probability\ of\ Sale} = e^{f(all)}$$

where

$$\begin{aligned} f(all) = {} & -1.8033 - 1.4758\ (Price\ Ratio) - 0.1940\ (Brand\ Bool)(Price\ Ratio) + 0.0577\ (Price\ Ratio)^2 \\ & + 0.2981\ (Star\ Rating\ Centred) + 0.1686\ (Review\ Score\ Centred) + 0.3927\ (Brand\ Bool) \\ & + 0.1017\ (Location\ Score1\ Centred) + 0.8330\ (Historical\ Price\ Ratio) \\ & - 0.1517\ (Historical\ Price\ Ratio)^2 - 0.1100\ (Position) + 0.1435\ (Promotion\ Flag) \end{aligned}$$

To analyse the effects of price ratio when other variables are kept constant, we rewrite *f(all)* into two constituent parts.

$$f(all) = f(Price\ Ratio) + f(others)$$

$$f(Price\ Ratio) = -1.4758\ (Price\ Ratio) - 0.1940\ (Brand\ Bool)(Price\ Ratio) + 0.0577\ (Price\ Ratio)^2$$

$$\begin{aligned} f(others) = {} & -1.8033 + 0.2981\ (Star\ Rating\ Centred) + 0.1686\ (Review\ Score\ Centred) + 0.3927\ (Brand\ Bool) \\ & + 0.1017\ (Location\ Score1\ Centred) + 0.8330\ (Historical\ Price\ Ratio) \\ & - 0.1517\ (Historical\ Price\ Ratio)^2 - 0.1100\ (Position) + 0.1435\ (Promotion\ Flag) \end{aligned}$$

Then we can rewrite the estimated odds of sale in:

$$Odds\ of\ Sale = e^{f(Price\ Ratio) + f(others)} = e^{f(Price\ Ratio)} e^{f(others)}$$

From the above, we can see that the part $e^{f(Price\ Ratio)}$ is the multiplication effects of "Price Ratio" on Odds of Sale, keeping all other variables constant. Recall that $e^{f(Price\ Ratio)}$ can be expanded into:

$$e^{f(Price\ Ratio)} = e^{-1.4758\ (Price\ Ratio) - 0.1940\ (Brand\ Bool)(Price\ Ratio) + 0.0577\ (Price\ Ratio)^2}$$

Since we are interested in the probability of sale for a hotel chain, brand_bool is equal to 1 and the equation can be written as:

$$e^{f(Price\ Ratio)} = e^{-1.6698\ (Price\ Ratio)+0.0577\ (Price\ Ratio)^2}$$

This is the marginal odds of "Price Ratio" (i.e. multiplication effects on Odds of Sale) and is depicted in the below figure.

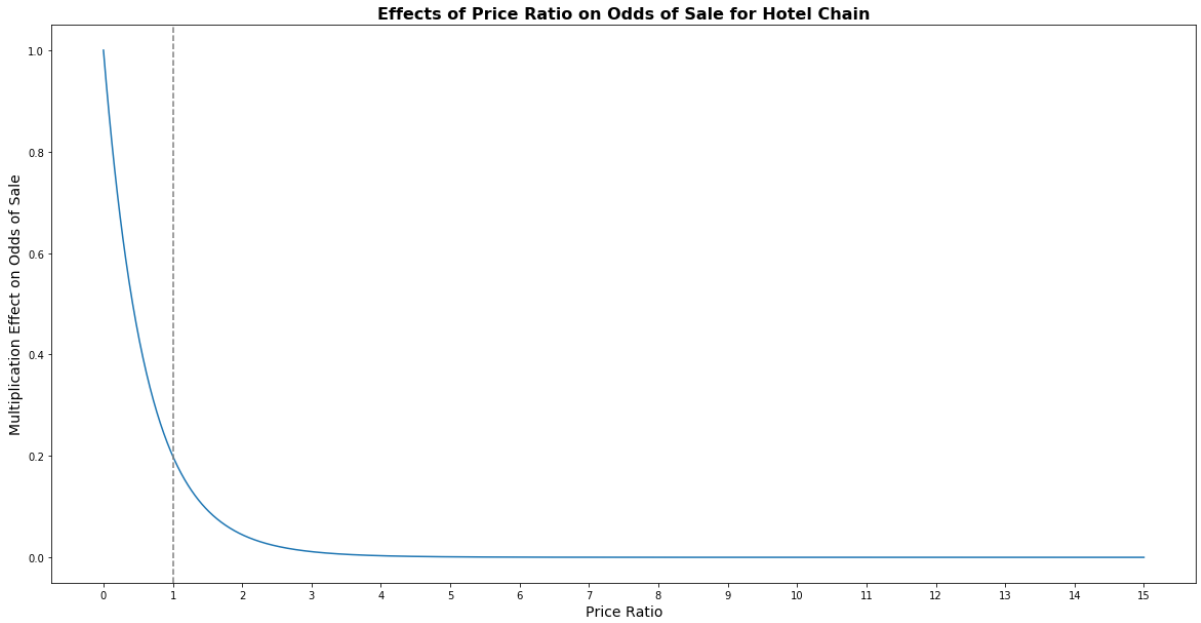

Figure 15 Multiplication Effects of Price Ratio on Odds of Sale for Hotel Chain (brand_bool = 1)

The vertical line indicates the position where Price Ratio = 1 (i.e. hotel price is equal to the average price of competitors). As we can see, the hotel chain should charge less than the average to improve the odds of sale. The lesser the hotel chain charge, the higher the odds of sale (and hence higher probability of sale).

In summary, the hotel chain should charge as low as possible compared to the average price of competitors. However, the hotel chain should also charge a rate that is profitable to the hotel. It is not financially prudent to charge a very low hotel rate to improve probability of sale but incur losses to the hotel chain. Furthermore, the hotel chain should not charge too low rate on Expedia to avoid undermining its own loyalty programme (for booking through its own website).

Optimal Price to Maximise Expected Revenue

To maximise the expected revenue, we would need the find the maximum of the function:

$$Function\ to\ Maximise = Probability\ of\ Sale * Price$$

Since $Price\ Ratio = \frac{Price}{Average\ Competitor\ Price}$ and the average price of competitors is constant for a given search results, we can maximise the following function instead.

$$Function\ to\ Maximise = Probability\ of\ Sale * Price\ Ratio$$

In previous section, we have been focusing on "Odds of Sale". To maximise the expected revenue, we would need the probability of sale instead. In logistic regression, the estimated probability of sale is:

$$Probability\ of\ Sale = \frac{e^{f(all)}}{1 + e^{f(all)}}$$

To reduce the probability of sale into a function of price ratio, we need the values of all other independent variables. Without loss of generality, we assume the following values of the other independent variables for our hotel chain:

Star Rating (Centred)	Review Score (Centred)	Brand Bool	Location Score1 (Centred)	Historical Price Ratio	Position	Promotion Flag
0.0	0.0	1	0.0	1.0	3	0

Table 14 Assumed Values of Independent Variables

With the above values, *f(all)* and Probability of Sale become:

$$f(all) = -1.8033 - 1.4758\ (Price\ Ratio) - 0.1940\ (1)(Price\ Ratio) + 0.0577\ (Price\ Ratio)^2 + 0.2981\ (0) + 0.1686\ (0) + 0.3927\ (1) + 0.1017\ (0) + 0.8330\ (1) - 0.1517\ (1)^2 - 0.1100\ (3) + 0.1435\ (0)$$

$$f(all) = -1.0593 - 1.6698\ (Price\ Ratio) + 0.0577\ (Price\ Ratio)^2$$

$$Probability\ of\ Sale = \frac{e^{-1.0593-1.6698\ (Price\ Ratio)+0.0577\ (Price\ Ratio)^2}}{1 + e^{-1.0593-1.6698\ (Price\ Ratio)+0.0577\ (Price\ Ratio)^2}}$$

$$Function\ to\ Maximise = \left(\frac{e^{-1.0593-1.6698\ (Price\ Ratio)+0.0577\ (Price\ Ratio)^2}}{1 + e^{-1.0593-1.6698\ (Price\ Ratio)+0.0577\ (Price\ Ratio)^2}}\right) * (Price\ Ratio)$$

Using scipy optimizer in Python, we can find the optimal price ratio that optimizes this function. The optimal price ratio is found to be 0.6991. In other words, to optimise the expected revenue, the hotelier should charge around 30% less than the average price of competitors.

The following graph plots expected revenue as a function of price ratio. We can see that the optimal price ratio is indeed around 0.6991 (the grey dashed line).

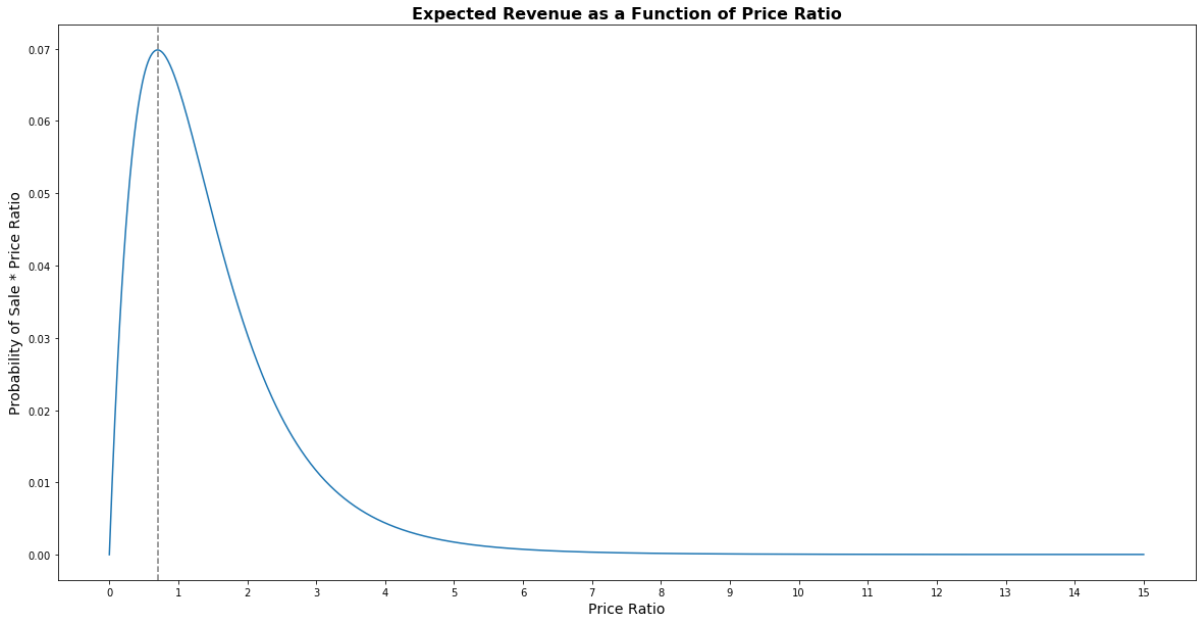

Figure 16 Expected Revenue as a Function of Price Ratio

References

Kaggle. (2017) *Personalize Expedia Hotel Searches – ICDM 2013*. Available from: https://www.kaggle.com/c/expedia-personalized-sort [Accessed 31st May 2017]

Scikit-Learn Developers. (2016) *Clustering*. Available from: http://scikit-learn.org/stable/modules/clustering.html [Accessed 13th June 2017]

Knowledgetack. (2016) *Two-sample Hypothesis Testing in Python with StatsModels*. Available from: http://knowledgetack.com/python/statsmodels/two-sample-hypothesis-testing-in-python-with-statsmodels/ [Accessed 13th June 2017]

Appendices

Logistic Regression Results of Four Models (Part 1, Question 2)

	Dependent variable:
	booked_bool
Constant	1.6622*** (0.0289)
geo_diverse	−0.0626*** (0.0016)
price_quality_diverse	−0.0001*** (0.00004)
visitor_location_id	0.0010*** (0.0001)
prop_country_id	−0.0003*** (0.0001)
length_of_stay	−0.2198*** (0.0031)
booking_window	−0.0054*** (0.0001)
search_adult_count	−0.1612*** (0.0064)
search_child_count	−0.0080 (0.0067)
room_count	0.3282*** (0.0144)
saturday_bool	−0.1741*** (0.0113)
Observations	182,127
Log Likelihood	−105,410.5000
Akaike Inf. Crit.	210,843.0000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 15 Part 1 Question 2 – Model 1 Logistic Regression Results

	Dependent variable:
	booked_bool
Constant	1.5545*** (0.0277)
geo_diverse	−0.0364*** (0.0012)
price_diverse	−0.00002*** (0.00001)
quality_diverse	0.0515*** (0.0136)
visitor_location_id	0.0011*** (0.0001)
prop_country_id	0.00002 (0.0001)
length_of_stay	−0.2210*** (0.0026)
booking_window	−0.0053*** (0.0001)
search_adult_count	−0.1686*** (0.0057)
search_child_count	−0.0053 (0.0057)
room_count	0.2867*** (0.0133)
saturday_bool	−0.1573*** (0.0094)
Observations	266,385
Log Likelihood	−152,778.2000
Akaike Inf. Crit.	305,580.3000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 16 Part 1 Question 2 – Model 2 Logistic Regression Results

	Dependent variable:
	no_click
Constant	1.1204*** (0.0073)
geo_diverse	−0.0008** (0.0004)
price_quality_diverse	−0.00001 (0.00001)
visitor_location_id	−0.00001 (0.00003)
prop_country_id	−0.0001*** (0.00003)
length_of_stay	0.0049*** (0.0007)
booking_window	−0.0001** (0.00003)
search_adult_count	0.0223*** (0.0016)
search_child_count	−0.0086*** (0.0018)
room_count	−0.0352*** (0.0032)
saturday_bool	0.0052* (0.0029)
Observations	182,127
R^2	0.0018
Adjusted R^2	0.0018
Residual Std. Error	0.5825 (df = 182116)
F Statistic	32.9516*** (df = 10; 182116)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 17 Part 1 Question 2 – Model 3 Logistic Regression Results

	Dependent variable:
	no_click
Constant	1.1444*** (0.0066)
geo_diverse	−0.0014*** (0.0003)
price_diverse	−0.000001 (0.000001)
quality_diverse	−0.0373*** (0.0033)
visitor_location_id	−0.000002 (0.00002)
prop_country_id	−0.0001*** (0.00002)
length_of_stay	0.0048*** (0.0006)
booking_window	−0.0001*** (0.00002)
search_adult_count	0.0228*** (0.0014)
search_child_count	−0.0079*** (0.0014)
room_count	−0.0353*** (0.0028)
saturday_bool	0.0055** (0.0022)
Observations	266,385
R^2	0.0023
Adjusted R^2	0.0022
Residual Std. Error	0.5484 (df = 266373)
F Statistic	55.6091*** (df = 11; 266373)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 18 Part 1 Question 2 – Model 4 Logistic Regression Results

Supplementary Figures (Part 1, Question 6)

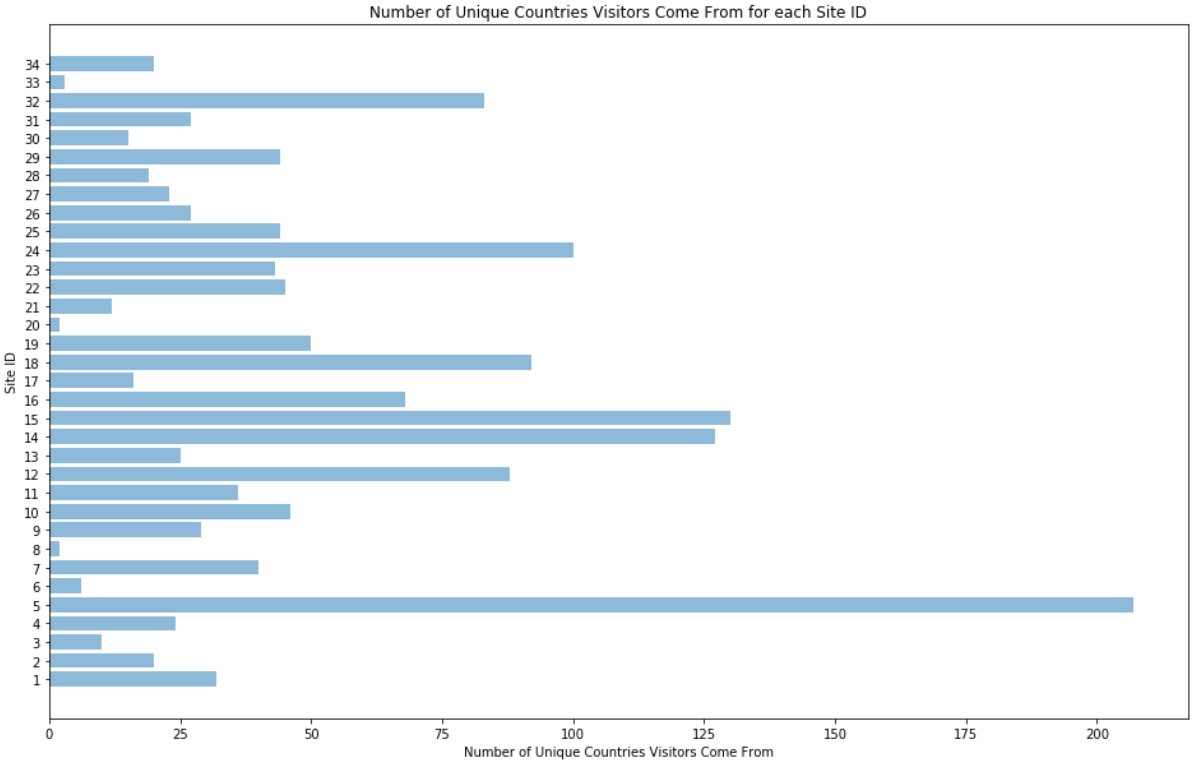

Figure 17 Number of Unique Countries of the Visitors to different Expedia Sites

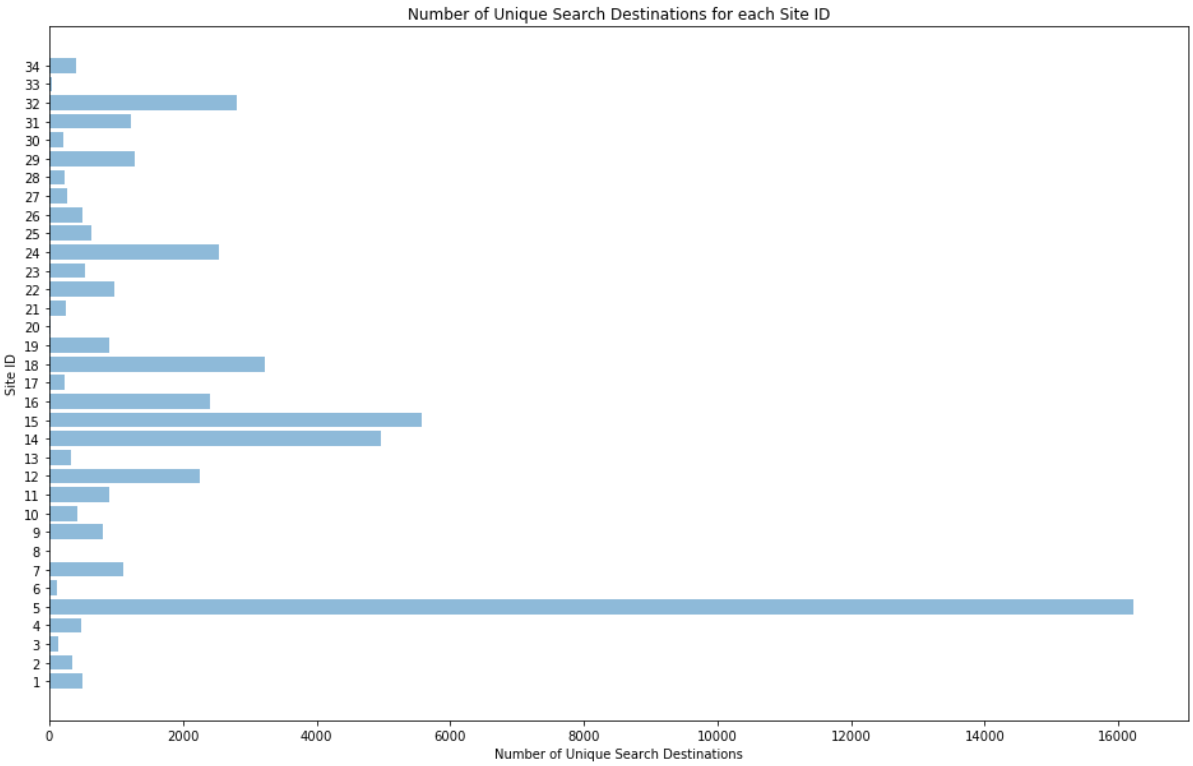

Figure 18 Number of Unique Search Destinations for different Expedia Sites

Reasoning Behind Independent Variable Selection (Part 2b)

Independent Variables	Derivation / Description
Price Ratio	Formula Hotel's Price / Average Price of Competing Hotels Reasoning • Each customer searches for different price range and hence the price difference (between hotels) in different searches can be vastly different in absolute dollar terms • To address this, the ratio is used to normalise the difference in percentage terms. For example, we can examine the effects when the hotel pricing is 10% higher than the average (instead of using absolute dollar difference)
Property Brand Bool : Price Ratio	*This is an interaction between the brand boolean variable and the price ratio defined above* Formula prop_brand_bool * price ratio Reasoning • Interaction term is added to examine if there is any different in price-sensitivity when the hotel type is different (i.e. chain vs independent) • Hotel chains listed on Expedia might be more price-sensitive since hotel chains usually offer loyalty programs at their own websites • We are interested in the scenario where prop_brand_bool = 1 since we are answering from hotel chain perspective
(Price Ratio)^2	Formula (Price Ratio)^2 Reasoning • This quadratic term is added to investigate the nonlinear relationships between price and the probability of sale • The probability of sale might not increase/decrease monotonically with respect to Price Ratio. The gradient might reverse its direction after a certain value and this variable is to examine this
Property Star Rating (Centred)	Formula Hotel's Star Rating – Average Star Rating of Competing Hotels Reasoning • Similar to the reasons described in "Price Ratio", each customer search may focus on different range of hotels, hence the average star ratings for each search can be vastly different • The probability of sale depends on how luxurious is our hotel in relative to the competing hotels (which show up in the same search). Therefore, we "centre" the star rating by subtracting average star rating
Property Review Score (Centred)	Formula Hotel's Review Score – Average Review Score of Competing Hotels Reasoning • Similar to the reasons described in "Property Star Rating (Centred)", different range of property review scores may show up in different searches, depending on the search filter a user set (e.g. budget) • The probability of sale depends on how well-received is our hotel in relative to the competing hotels (which show up in the same search). Therefore, we "centre" the review score by subtracting average review score
Property Brand Bool	Formula Brand boolean indicating the type of hotel (1 = hotel chain, 0 = independent hotel)

	Reasoning • Different hotel types may have different probability of being booked, hence this boolean is included to control for this factor
Property Location Score1 Centred	Formula Hotel's Location Score1 – Average Location Score1 of Competing Hotels Reasoning • Similar to the reasons described in "Property Star Rating (Centred)", different searches may show vastly different range of property location scores, depending on the user preference. By subtracting the average location score, we focus on the relative score difference in the same search • Since "prop_location_score1" and "prop_location_score2" both measure the location desirability (and hence highly correlated), only "prop_location_score1" is selected in the regression since a lot of the hotels do not have "prop_location_score2" (around 20%)
Historical Price Ratio	*Note: In the Expedia dataset, the historical price is given in logarithm form. Exponent is taken to obtain historical price in decimal form.* Formula Hotel's Mean Price in Last Period / Average Mean Price of Competing Hotels in Last Period Reasoning • This acts as a control variable in the regression. A hotel room, which was sold in higher mean price (in last period) compared with the competitors, may indicate higher popularity. Therefore, this variable is used to control for the popularity factor in the regression • Similar to the reasons described in "Price Ratio", we are interested in the percentage difference in historical price. Therefore ratio is used
(Historical Price Ratio)^2	Formula (Historical Price Ratio)^2 Reasoning • Similar to the reasons described before, the relationships between probability of sale and historical price ratio may not be linearly increasing / decreasing. Thus, this quadratic term is added as regressor
Position	Formula Hotel's position on Expedia's search results page Reasoning • Hotel that is placed higher up at the search results page might get higher probability of being acted upon, since user is less likely to scroll down the page and examines all the hotel listing. Therefore, it is important that we control for this factor in the regression
Promotion Flag	Formula Binary variable indicating whether the hotel has a sales promotion Reasoning • A user might be more likely to book a hotel that has sales promotion if it is a good bargain. This is an important variable to control for in the regression

Table 19 Independent Variables Used in Logistic Regression