

# Prediction Analysis on Hotel booking cancellation

Logistic regression on a Hotel's marketing data



1. Introduction and Overview .....	2
2. Marketing Data summarised .....	2
3. Finding probability of Booking cancellation using Logistic Regression Analysis .....	7
What is Logistic regression? .....	7
Performing Logistic Regression.....	8
Applying Logistic Regression to our use case .....	8
Steps for Logistic regression: .....	8
4. Evaluation of Outcome .....	9
5. Conclusion and recommendations .....	11
6. References .....	11
Appendix.....	12

## 1. Introduction and Overview

This report focusses on predicting the booking cancellation of 2 hotels based on their historical data. The master dataset used for this report consists of more than 111,000 records of customer bookings combined for both the hotels from July 2015 to August 2017. One of the hotels is a plush Resort whereas the other one is a City based hotel (Antonio, de Almeida and Nunes, 2019). Since these are hotels with actual data relevant to real people, sensitive information such as the identification details of the Hotels and the customers is omitted as per the GDPR guidelines (Greengard, 2018). The type of database in this case is a 'Structured' data as it is extracted from the Hotel's SQL databases and is highly organised into tables and columns (Gandomi and Haider 2015).

The booking system of these hotels allow 3 types of bookings. They include direct booking, corporate booking, and booking by a Travel agent (Antonio, de Almeida and Nunes, 2019). As a part of marketing and expansion plan, Hotel operators have started operating websites as well as tie ups with OTAs i.e. Online Travel Agents. However, it is clear from various reports that this approach did not help increase their profits/revenues. An example of this is the website development investment from the Four Seasons hotel who invested 18 million USD for a marketing and booking, but it increased only 2% of revenue in a duration of 5 years (Liu and Zhang 2014). Also, the revenue from online bookings contributed only 12% of the total revenue (Fox 2020).

The scope of the report is to summarise and understand the different factors for predicting the cancellation of bookings for the hotels. The report will first carry out Descriptive statistics for summarising the variables followed by Logistic regression to find probability of booking cancellation in the future.

## 2. Marketing Data summarised

This section summarises the important variables involved, their types, and significance with descriptive statistics. As mentioned in Appendix 1, there are more than 30 variables as part of the master data. Out of those, only relevant variables were chosen as per studies to see their impact.

### a. is\_cancelled

This variable shows the number of bookings which were cancelled and not cancelled after being placed. This is a Boolean type of a variable

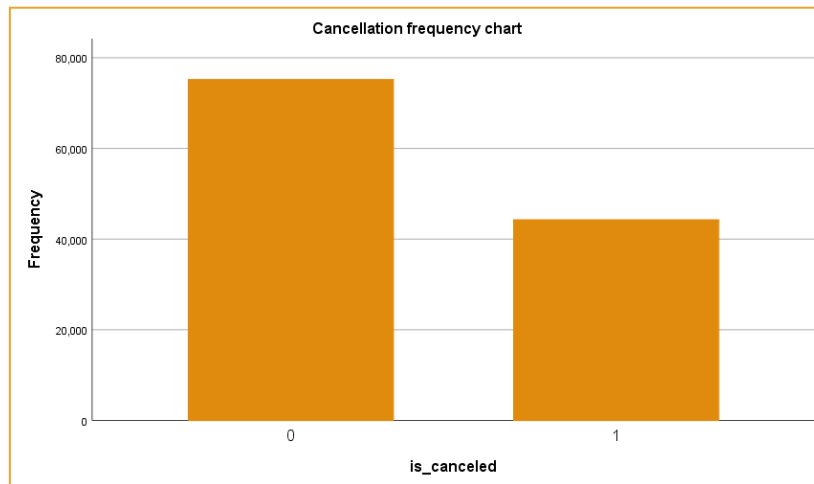


Fig. 2.a.1: Frequency of booking cancellation for hotel rooms (Refer Appendix 1 for data)

As visible in Fig. 01 and 02 the number of bookings which were cancelled were almost 37% of the total bookings done which is high in terms of cancellation. This is a revenue diminishing factor for the hotels. Online offers and schemes such as 'Free cancellation' and 'Click to Cancel' have increased the hotel booking cancellation frequency (Falk and Vieru 2018).

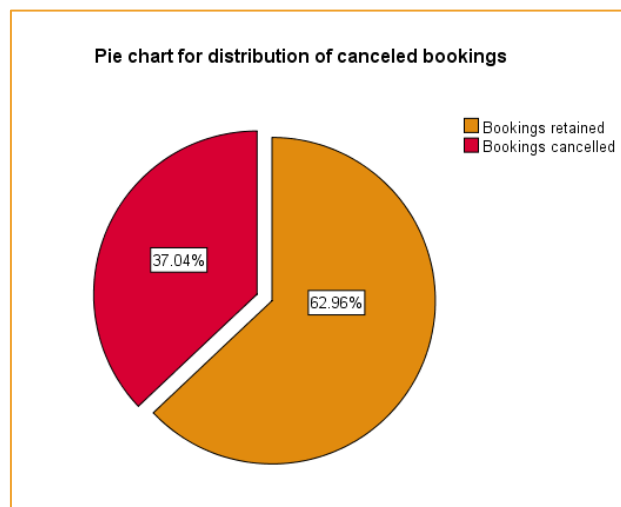


Fig. 2.a.2: Cancelled booking percentage (Refer appendix 1 for data)

b. **deposit\_type**

This variable suggests the type of booking deposit provided for booking a room in the hotel. This is a nominal type of variable.

The online bookings offered usually does not require deposit to be provided. OTAs provide 'No deposit' which is completely refundable (Delgado 2020). In a study, it was found that cancellation rates were highest for online booking (Falk and Vieru 2018). These were around 17%.

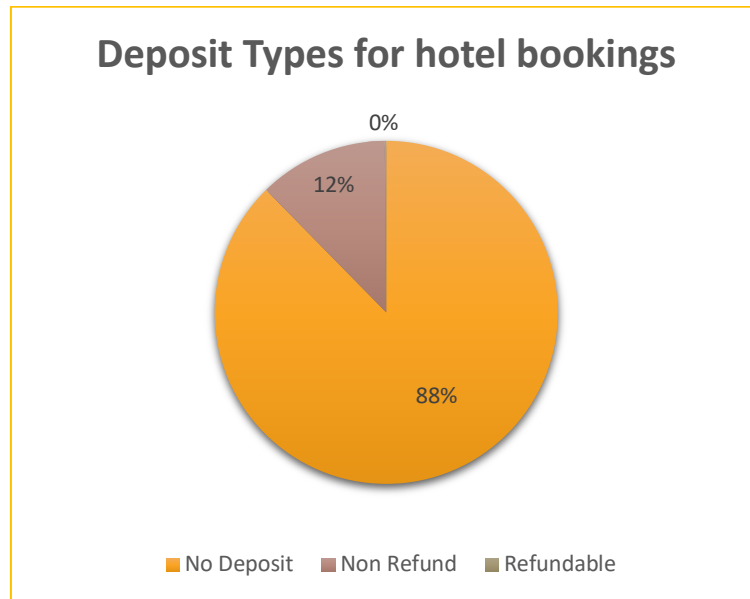


Fig. 2.b.1: Pie chart depicting Deposit types for Hotel bookings (refer Appendix 2.b)

As seen in the Fig 2.b.1, 87.6% of the bookings made did not require a deposit. Only 12.2 percent provided a non-refundable deposit in case the user cancels his booking.

c. **arrival\_data\_month**

The month in which the guests are due to arrive is an ordinal type of variable. It is observed that the number of cancellations is higher during the low demand months (Antonio et al. 2017). This is evident in Fig. 2.c.1 where months such as July and August, the number of bookings is very high, but the number of cancellations drops.

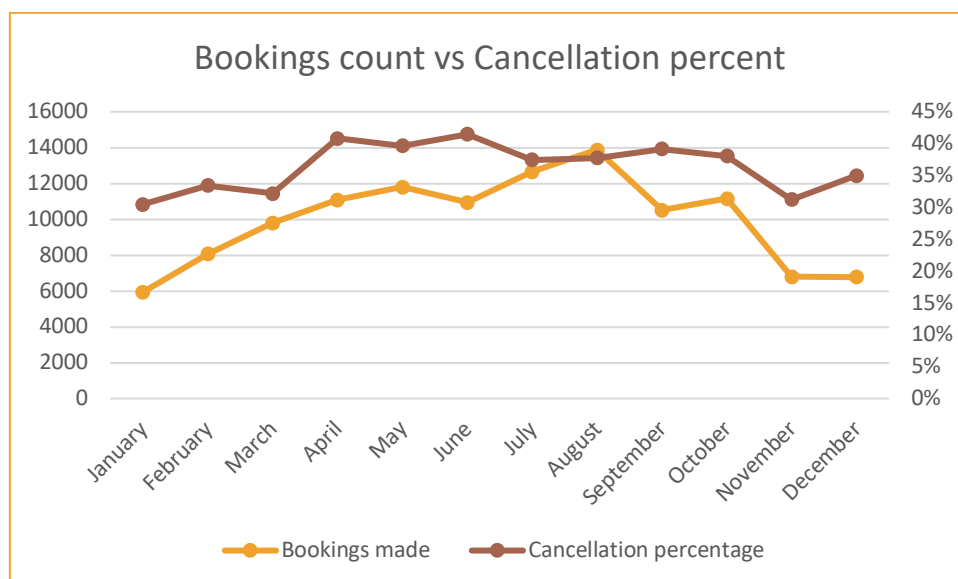


Fig 2.c.1: Comparison of Bookings made vs Bookings cancelled

d. **lead\_time**

This is a numerical integer variable which represents the count of days between booking date and the arrival date (Antonio et al. 2017). The likelihood of cancellation is higher in case of early bookings (Falk and Vieru 2018). Descriptive statistics is applied to this variable to gain better insight.

Mean	Median	Mode	Std. Deviation	Skewness	Kurtosis	Range	Minimum	Maximum
104	69.00	0	106.863	1.347	1.696	737	0	737

Table: 01 Descriptive statistics for lead time booking

Table 01 and Fig. 2.d.1 shows there is a huge difference between the values of Mean and Median suggesting the uneven spread of the lead time for booking. On the other hand, most bookings are done on the same date as arrival as given by the Mode. There is high skewness of 1.347 in the lead time data. The data is positively skewed as a significant number of values are present from 104 days to 737 days.

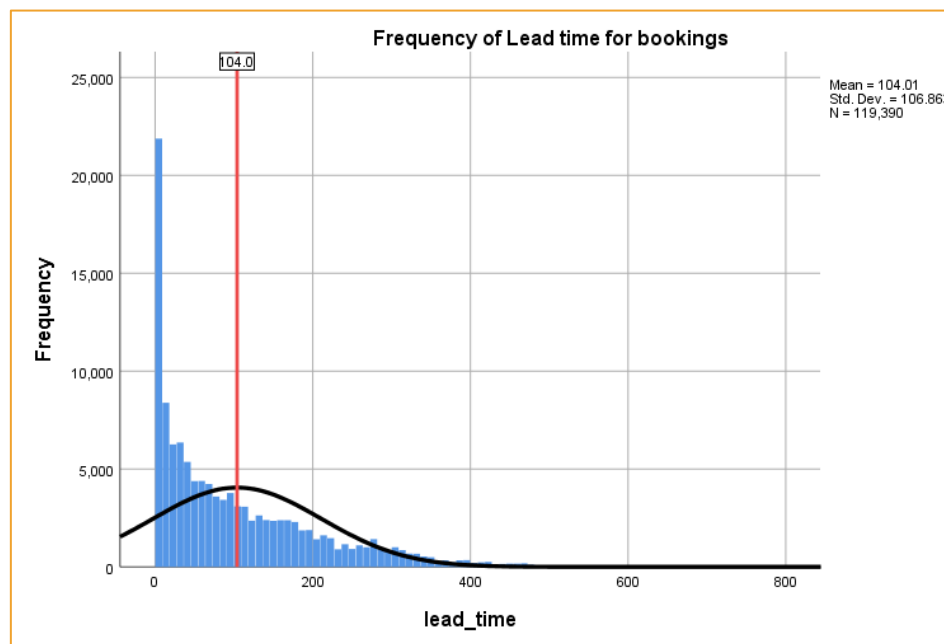


Fig 2.d.1. Descriptive statistics on lead booking time

e. **country**

This is a categorical variable which provides information about a person's country of origin of booking. This variable is important and affects booking cancellations (Antonio et al. 2017).

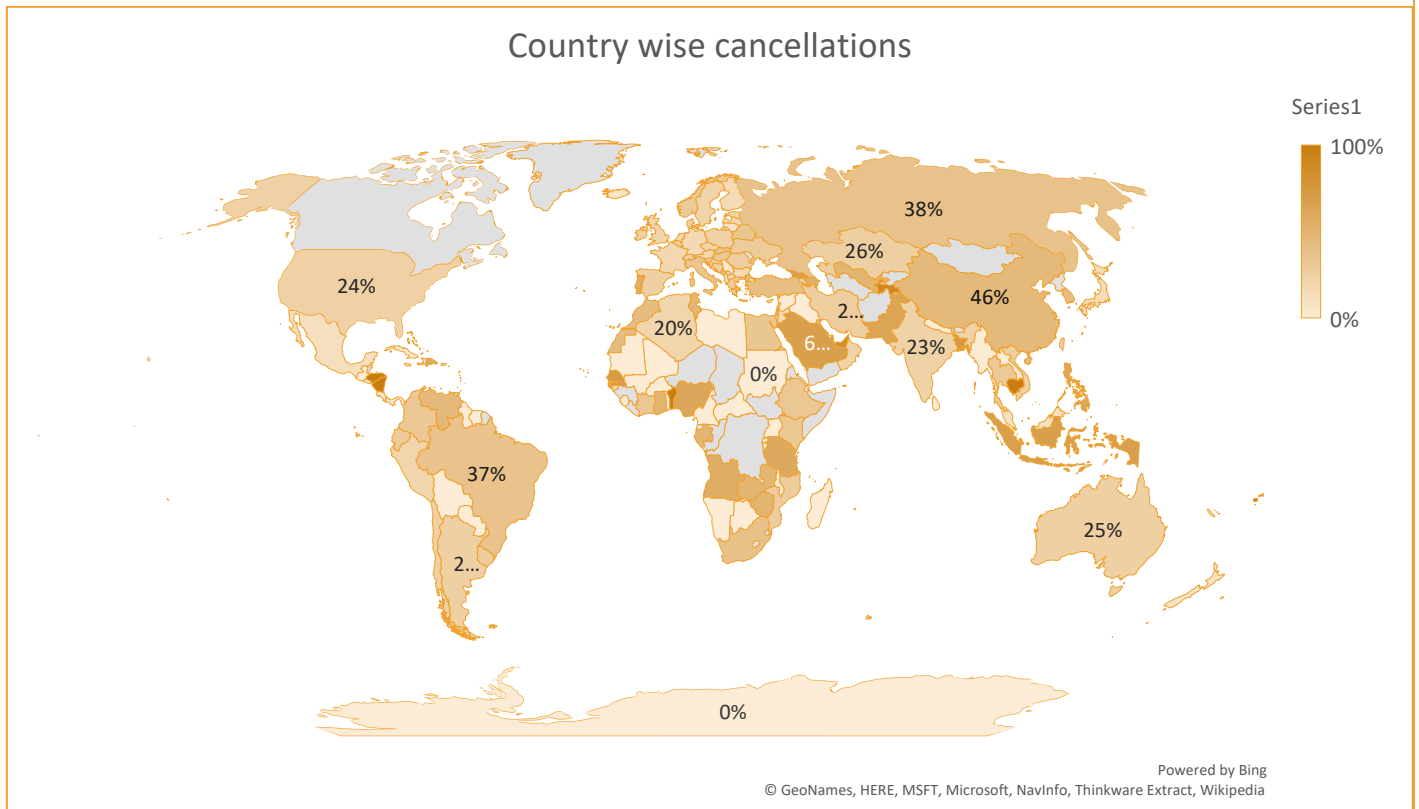


Fig 2.e.1. Map chart with percentage of country specific cancellations (refer Appendix 2.c)

Fig 2.e.1. shows cancellations of bookings based on origin country. This map chart is generated using Pivot table on the raw master data. The country names are conforming to the ISO 3166 country codes (Celko 2010). Highest number of bookings are done from Republic of Portugal with Cancellations as high as 57%. Countries such as Saudi Arabia, United Arab Emirates, Hong Kong, Macao have cancellation rates higher than 75%.

### 3. Finding probability of Booking cancellation using Logistic Regression Analysis

The motive of the report is to find the probability of users cancelling their room bookings for the case of the two hotels. To find out this probability the best method to use is Logistic regression.

#### What is Logistic regression?

This is a methodology to find a model to represent the probability of an event occurring depending on values of independent variables. This helps to provide information of a series of variables on a binary response variable. This binary response is called a dependent variable and in our case it is the booking cancellation. As depicted in Fig. 3.a, the model helps to identify if the output is '1' or '0' based on the provided input.



Fig. 3.a Logistic regression model high level process

Logistic regression helps to find this probability of cancellation based on the data/variables related to cancellation data. As per Grégoire (2014), the intention of Logistic regression is to find the relationship between the probability of an event occurring and a set of related variables. This methodology would help the hotel to find out if a particular combination of variables may lead to bookings being cancelled.

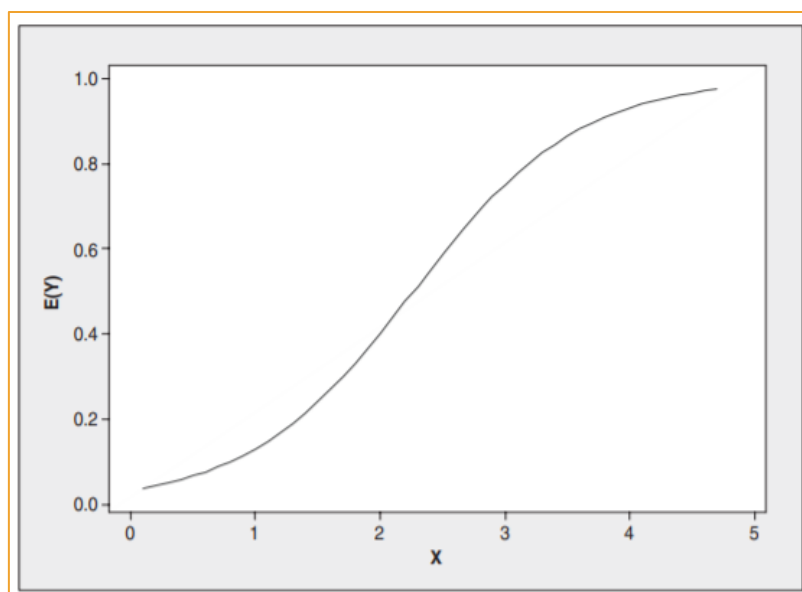




Fig. 3.b Sigmoid curve for a binomial logistic regression

The Fig. 3.b shows an S curve which is applicable for a 'Y' which is a binary random variable – as in our case. The function used here is also called as a 'logit' function.

$$p(x) = (\exp(\beta_0 + \beta_1 X)) / (1 + \exp(\beta_0 + \beta_1 X))$$

X is the independent known variable whereas  $\beta_0$  is the intercept. 'exp' stands for the exponential value of the chosen variable. This being a probability, the value would always remain between 0 and 1 for a range of different values of X. This is the standard model for a Binomial Logistic regression ("Logistic Regression" 2018).

The tool used for logistic regression in this case is Microsoft excel (Microsoft Excel Spreadsheet Software 2020)

### Performing Logistic Regression

In our case the binary dependent variable i.e. the variable whose probability is to be found is Booking cancellation. This means that the output is going to either a '1' i.e. Booking cancelled or a '0' i.e. Booking not cancelled.

We have used the Maximum likelihood estimation method which help to find out the best parameters/coefficient for the model. In other words, the Maximum Likelihood Estimation is a method that determines the values for the parameters. The parameter values are found such that they maximise the likelihood that the process described by the model produced the data that were actually observed.

### Applying Logistic Regression to our use case

In our case, there are two independent variables which can be used to find the probability. This requires using a different variant of the function.

$$p(x) = (\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)) / (1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)) \quad \text{----- formula (1)}$$

Where X1 and X2 are the two independent variables. The calculation will be done manually using Microsoft Excel with 'Solver' add in integrated within the tool.

The best variables to choose for this use case are Lead time and the Average Daily Rate for the Hotels (Antonio et al. 2017). Both these variables are numerical and are able to provide a better forecast of probability.

### Steps for Logistic regression:

1. Add the Solver add on for Microsoft excel (Load The Solver Add-In In Excel 2020)

2. 1000 records from the master dataset are chosen randomly to apply Logistic regression
3. These records are then pasted into Microsoft excel for analysis
4. The data is first cleaned to remove any record with missing values
5. For creating the model, this manual process is done stepwise
  - a. The equation ' $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ ' is first calculated by assuming the values for  $X_1$  and  $X_2$  as 0.01.
  - b. Then Solver add in is invoked to run on the input variables (i.e.  $X_1$  and  $X_2$ ) to find out the best possible values for  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$ . The values to be found is non-linear so solving method selected is non-linear.
  - c. Solver runs to calculate the best possible values based on the given inputs. This generated the equation  $-2.3675 + 0.0057 * X_1 + 0.00958 * X_2$
  - d. Once generated, this is then used to calculate all the values with both the variables  $X_1$  and  $X_2$  to apply the Excel Exponential function (EXP Function Excel 2020)
  - e. Once the EXP values are calculated, the probability is then determined using the formula (1) for all the values. Refer Appendix 3.a and Fig 4.b

#### 4. Evaluation of Outcome

Once all the values are calculated, the S curve is plotted for probability of the booking cancellation. Fig. 4.a shows the probability of booking cancellation plotted. This is called as sigmoid curve and the values on Y axis always lie between 0 and 1 (Su and Raghavarao 2013).

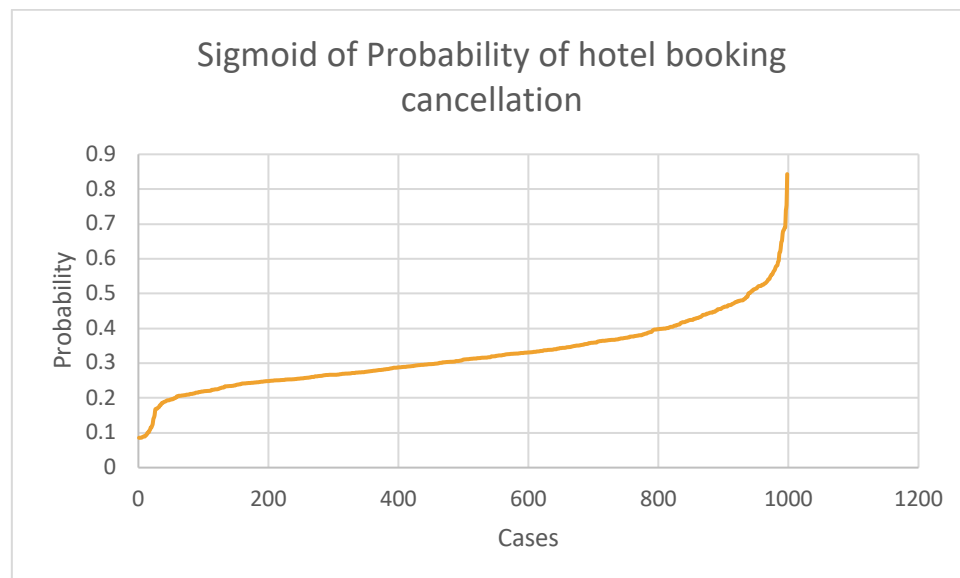


Fig. 4.a S curve for probability of Hotel booking cancellation

The Fig. 4.a explains us that with the increase of lead time and the average daily rate, the probability of the hotel guests to cancel his booking increases as well. The Average daily rate which is calculated by sum of all the lodging transactions by the total number of staying nights, has more impact on the user's booking. This is explained as below. As shown in 4.b, the  $P(x)$  and  $1-P(x)$  represent the

probability of success and the probability of failure. In this case  $P(x)$  represents the probability of booking getting cancelled and  $1-P(x)$  represents the probability of the booking not getting cancelled.

	B	C	D	E	F	G	H	I	J	K	L	M
	is_cancelled	lead_time (X1)	adr (X2)	Forecast	Exponential Linear regression	P(x)	1-P(x)	Likelihood	Logarithm of likelihood (ln)	Sum	Intercept	
1	0	7	75	=SMIS1+SMIS2*C2 + SMIS3*D2	=EXP(E2)	=F2/(1+F2)	=1-G2	=IF(B2=1,G2,H2)	=LN(I2)	=SUM(J2:J999)	Slope 1	-2.36759484605054
2	0	13	75	=SMIS1+SMIS2*C3 + SMIS3*D3	=EXP(E3)	=F3/(1+F3)	=1-G3	=IF(B3=1,G3,H3)	=LN(I3)		Slope 2	0.00579763942738713
3	0	14	98	=SMIS1+SMIS2*C4 + SMIS3*D4	=EXP(E4)	=F4/(1+F4)	=1-G4	=IF(B4=1,G4,H4)	=LN(I4)			0.00958365037624077
4	0	14	98	=SMIS1+SMIS2*C5 + SMIS3*D5	=EXP(E5)	=F5/(1+F5)	=1-G5	=IF(B5=1,G5,H5)	=LN(I5)			
5	0	0	107	=SMIS1+SMIS2*C6 + SMIS3*D6	=EXP(E6)	=F6/(1+F6)	=1-G6	=IF(B6=1,G6,H6)	=LN(I6)			
6	0	9	103	=SMIS1+SMIS2*C7 + SMIS3*D7	=EXP(E7)	=F7/(1+F7)	=1-G7	=IF(B7=1,G7,H7)	=LN(I7)			
7	1	85	82	=SMIS1+SMIS2*C8 + SMIS3*D8	=EXP(E8)	=F8/(1+F8)	=1-G8	=IF(B8=1,G8,H8)	=LN(I8)			
8	1	75	105.5	=SMIS1+SMIS2*C9 + SMIS3*D9	=EXP(E9)	=F9/(1+F9)	=1-G9	=IF(B9=1,G9,H9)	=LN(I9)			
9	1	23	123	=SMIS1+SMIS2*C10 + SMIS3*D10	=EXP(E10)	=F10/(1+F10)	=1-G10	=IF(B10=1,G10,H10)	=LN(I10)			
10	1	35	145	=SMIS1+SMIS2*C11 + SMIS3*D11	=EXP(E11)	=F11/(1+F11)	=1-G11	=IF(B11=1,G11,H11)	=LN(I11)			
11	0	68	97	=SMIS1+SMIS2*C12 + SMIS3*D12	=EXP(E12)	=F12/(1+F12)	=1-G12	=IF(B12=1,G12,H12)	=LN(I12)			
12	0	18	154.77	=SMIS1+SMIS2*C13 + SMIS3*D13	=EXP(E13)	=F13/(1+F13)	=1-G13	=IF(B13=1,G13,H13)	=LN(I13)			
13	0	37	94.71	=SMIS1+SMIS2*C14 + SMIS3*D14	=EXP(E14)	=F14/(1+F14)	=1-G14	=IF(B14=1,G14,H14)	=LN(I14)			
14	0	68	97	=SMIS1+SMIS2*C15 + SMIS3*D15	=EXP(E15)	=F15/(1+F15)	=1-G15	=IF(B15=1,G15,H15)	=LN(I15)			
15	0	37	97.5	=SMIS1+SMIS2*C16 + SMIS3*D16	=EXP(E16)	=F16/(1+F16)	=1-G16	=IF(B16=1,G16,H16)	=LN(I16)			
16	0	12	88.2	=SMIS1+SMIS2*C17 + SMIS3*D17	=EXP(E17)	=F17/(1+F17)	=1-G17	=IF(B17=1,G17,H17)	=LN(I17)			
17	0	0	107.42	=SMIS1+SMIS2*C18 + SMIS3*D18	=EXP(E18)	=F18/(1+F18)	=1-G18	=IF(B18=1,G18,H18)	=LN(I18)			
18	0	7	153	=SMIS1+SMIS2*C19 + SMIS3*D19	=EXP(E19)	=F19/(1+F19)	=1-G19	=IF(B19=1,G19,H19)	=LN(I19)			
19	0	37	97.29	=SMIS1+SMIS2*C20 + SMIS3*D20	=EXP(E20)	=F20/(1+F20)	=1-G20	=IF(B20=1,G20,H20)	=LN(I20)			
20	0	72	84.67	=SMIS1+SMIS2*C21 + SMIS3*D21	=EXP(E21)	=F21/(1+F21)	=1-G21	=IF(B21=1,G21,H21)	=LN(I21)			
21	0	72	84.67	=SMIS1+SMIS2*C22 + SMIS3*D22	=EXP(E22)	=F22/(1+F22)	=1-G22	=IF(B22=1,G22,H22)	=LN(I22)			
22	0	72	99.67	=SMIS1+SMIS2*C23 + SMIS3*D23	=EXP(E23)	=F23/(1+F23)	=1-G23	=IF(B23=1,G23,H23)	=LN(I23)			
23	0	127	94.55	=SMIS1+SMIS2*C24 + SMIS3*D24	=EXP(E24)	=F24/(1+F24)	=1-G24	=IF(B24=1,G24,H24)	=LN(I24)			
24	0	78	63.6	=SMIS1+SMIS2*C25 + SMIS3*D25	=EXP(E25)	=F25/(1+F25)	=1-G25	=IF(B25=1,G25,H25)	=LN(I25)			
25	0	48	79.5	=SMIS1+SMIS2*C26 + SMIS3*D26	=EXP(E26)	=F26/(1+F26)	=1-G26	=IF(B26=1,G26,H26)	=LN(I26)			
26	1	60	107	=SMIS1+SMIS2*C27 + SMIS3*D27	=EXP(E27)	=F27/(1+F27)	=1-G27	=IF(B27=1,G27,H27)	=LN(I27)			
27	1	77	94	=SMIS1+SMIS2*C28 + SMIS3*D28	=EXP(E28)	=F28/(1+F28)	=1-G28	=IF(B28=1,G28,H28)	=LN(I28)			
28												

Fig. 4.b: Formulae used for the manual calculation of Probability of Booking cancellation

In Logistic regression, to estimate an unknown probability for a linear combination of independent variables, we link the independent variables together using the Bernoulli distribution.

The values of coefficients which are  $\beta_1$  and  $\beta_2$  is 0.0057 and 0.00958 respectively.  $\beta_1$  is associated with Lead time of booking whereas  $\beta_2$  is associated with the Average Daily rate. This means that for every 1 unit increase in lead time (day) the probability of cancellation is increased by 0.5%. On the other hand, for every one unit increase in Average Daily rate of the hotel, the probability of the user cancelling his booking is increased by almost 1% (Hilbe 2017). This shows that the variable 'Average Daily rate' has a very high effect on the booking cancellation. Our calculations show that when the Lead time is increased by more than 200 days and the Average Daily rate has increased by more than 200, the probability of cancellation is almost 100%.

The accuracy of this model could be increased more by adding additional variables to the model. However, the current model is still convergent. We also have a Maximum Likelihood determination which has helped to find what is the maximum likelihood for the probability to occur. This helps us to find the best possible data for determining the probabilities. Now as we have plotted the curve and have the model ready, it can be further used to decrease the probability of a user cancelling his booking. Maximum likelihood tells us how probable the values at hand are (Hilbe 2017).

We need to understand that for the sake of this report only 2 independent variables were considered for calculation. However almost 30 variables were identified from the database which may have an impact on the user's booking cancellation behaviour. Section 2 depicted other variables such as country and booking deposit type which are also highly related as per the descriptive statistics carried out. Therefore, it is quite important to notice that these 2 variables considered for calculation do not represent the behaviour of customer booking completely.

Analysing both these variables, it is clear that the Hotels need to specifically concentrate and control their Average Daily Rate (ADR). This variable has a considerable effect on the consumer's behaviour.

It is an essential metric for Hotel performance. It represents the average rental income which is paid for each occupied room at a certain time and is calculated by dividing rooms's cost by the number of rooms sold (Oses et al. 2016). ADR can be estimated based on the cost of the hotel rooms online for a hotel. The coefficient value for ADR is very high i.e. 0.09 which is almost 0.1. This represents its high effectiveness in influencing the user's behaviour for booking cancellation.

The intercept value of -2.36 represents that the probability is quite low for the Model with these variables and coefficients. If the coefficients change their value to '0' then the probability of cancellation is reduced drastically.

## 5. Conclusion and recommendations

Higher 'Lead booking time' and the 'Average daily rate' are impacting the cancellation behaviour of the users. This is leading to reduced revenues and increased losses for the hotels. To avoid this, hotels are moving to strategies like overbooking (Dong and Ling 2015). However, this leads to unpleasant experiences to the hotels leading them to other options and more cancellations further down the road. But hotels are increasingly applying these strategies. The profit improvement from an overbooking strategy is around 4.20% whereas the profit improvement from online travel agents and websites is around 5.20%. This leads to the point that the Hotels should focus more on Online booking, phone booking and reduce cancellations (Dong and Ling 2015)

Online bookings allow the user to make a booking many days in advance to his arrival date. Also schemes and offers from such as 'No deposit booking' and 'Free cancellation' help the customers to book and later cancel the booking as there is no loss. Our results in Section 4 showed that the bookings made closer to the arrival date were less likely to be cancelled. So, it would be recommended to the hotels to keep a shorter Lead time window in order to reduce customer booking cancellations. It has been found that the customers booking through online travel agents have been offered a 12% higher price than the standard offline / phone booking rate (theguradian.com 2020). Such high prices can lead the customer to cancel the booking or not do it at all resulting in lower revenue for the hotels. It would be recommended for the hotels to track such behaviours, control them and encourage the customers to book with them directly using the hotel's online booking platform or phone banking to avoid unnecessary price hike and help the hotels increase their revenues (Antonio, N., Almeida, A. and Nunes, L. 2017). This was confirmed in the calculations in Section 4.

## 6. References

- Antonio, N., Almeida, A. and Nunes, L. (2017) "Predicting Hotel Booking Cancellations To Decrease Uncertainty And Increase Revenue". *Tourism & Management Studies* 13 (2), 25-39
- Antonio, N., de Almeida, A. and Nunes, L. (2019) "Hotel Booking Demand Datasets". *Data In Brief* 22, 41-49
- Celko, J. (2010) "ISO-3166 And Other Country Codes". *Joe Celko's Data, Measurements And Standards In SQL* 123-124
- Delgado, P. (2020) *Cancellations Shooting Up: Implications, Costs And How To Reduce Them* | [online] available from <<https://www.mirai.com/blog/cancellations-shooting-up-implications-costs-and-how-to-reduce-them/>> [28 February 2020]
- Dong, Y. and Ling, L. (2015) "Hotel Overbooking And Cooperation With Third-Party Websites". *Sustainability* 7 (9), 11696-11712

- Gandomi, A. and Haider, M. (2015) "Beyond The Hype: Big Data Concepts, Methods, And Analytics". *International Journal Of Information Management* 35 (2), 137-144
- Falk, M. and Vieru, M. (2018) "Modelling The Cancellation Behaviour Of Hotel Guests". *International Journal Of Contemporary Hospitality Management* 30 (10), 3100-3116
- Fox, L. (2020) *Four Seasons Unveils \$18 Million Website As Luxury Travel Grows* | Phocuswire [online] available from <<https://www.phocuswire.com/Four-Seasons-unveils-18-million-website-as-luxury-travel-grows>> [28 February 2020]
- Greengard, S. (2018) "Weighing The Impact Of GDPR". *Communications Of The ACM* 61 (11), 16-18
- Hilbe, J. (2017) *Logistic Regression Models*. London: Taylor & Francis Ltd
- Liu, J. and Zhang, E. (2014) "An Investigation Of Factors Affecting Customer Selection Of Online Hotel Booking Channels". *International Journal Of Hospitality Management* 39, 71-83
- "Logistic Regression" (2018) 312-384
- Microsoft Excel Spreadsheet Software* (2020) available from <<https://products.office.com/en-gb/excel>> [6 March 2020]
- Load The Solver Add-In In Excel (2020) available from <<https://support.office.com/en-gb/article/load-the-solver-add-in-in-excel-612926fc-d53b-46b4-872c-e24772f078ca>> [6 March 2020]
- Oses, N., Gerrikagoitia, J. and Alzua, A. (2016) "Modelling And Prediction Of A Destination'S Monthly Average Daily Rate And Occupancy Rate Based On Hotel Room Prices Offered Online". *Tourism Economics* 22 (6), 1380-1403
- theguardian.com (2020) *Travellers May Lose Out By Booking Hotels Online, Says Which?*. [online] available from <<https://www.theguardian.com/money/2020/mar/04/travellers-may-lose-out-by-booking-hotels-online-says-which>> [6 March 2020]
- Su, Y. and Raghavarao, D. (2013) "Minimal Plus One Point Designs For Testing Lack Of Fit For Some Sigmoid Curve Models". *Journal Of Biopharmaceutical Statistics* 23 (2), 281-293

## Appendix

### Appendix 1:

Entire dataset of Hotel bookings with all variables (Antonio, de Almeida and Nunes, 2019)



hotel\_bookings.csv

### Appendix 2.a:

#### Appendix 2.a:

Descriptive statistics for the variable 'deposit\_type' (Antonio, de Almeida and Nunes, 2019)

Booking amount type		
Deposit Type	Frequency	Percent
No Deposit	104641	87.6

Non Refund	14587	12.2
Refundable	162	0.1
Total	119390	100

**Appendix 2.b:**

Month wise bookings done (Antonio, de Almeida and Nunes, 2019)

Month	Frequency	Booking percentt	Cancellations	Cancellation percentage
January	5929	5%	1807	30%
February	8068	7%	2696	33%
March	9794	8%	3149	32%
April	11089	9%	4524	41%
May	11791	10%	4677	40%
June	10939	9%	4535	41%
July	12661	1%	4742	37%
August	13877	12%	5239	38%
September	10508	9%	4116	39%
October	11160	9%	4246	38%
November	6794	6%	2122	31%
December	6780	6%	2371	35%

**Appendix 2.c:**

Country Code	Percentage booking cancelled	Booking Retained	Booking Cancelled	Total Bookings done
BEN	100%	0	3	3
FJI	100%	0	1	1
GGY	100%	0	3	3
GLP	100%	0	2	2
HND	100%	0	1	1
IMN	100%	0	2	2
JEY	100%	0	8	8
KHM	100%	0	2	2
MYT	100%	0	2	2
NIC	100%	0	1	1
UMI	100%	0	1	1
VGB	100%	0	1	1
MAC	94%	1	15	16
HKG	90%	3	26	29
TJK	89%	1	8	9
ARE	84%	8	43	51
BHR	80%	1	4	5
FRO	80%	1	4	5
BGD	75%	3	9	12
MDV	75%	3	9	12
QAT	73%	4	11	15
SEN	73%	3	8	11
AND	71%	2	5	7

## 7013SSL – Applied Marketing Analytics

SAU	69%	15	33	48
IDN	69%	11	24	35
GEO	68%	7	15	22
PAK	64%	5	9	14
PHL	63%	15	25	40
NGA	62%	13	21	34
GIB	61%	7	11	18
TZA	60%	2	3	5
DOM	57%	6	8	14
PRT	57%	21071	27519	48590
AGO	57%	157	205	362
AZE	53%	8	9	17
CPV	50%	12	12	24
GAB	50%	2	2	4
GHA	50%	2	2	4
SYC	50%	1	1	2
UZB	50%	2	2	4
ZMB	50%	1	1	2
ZWE	50%	2	2	4
TUN	49%	20	19	39
CHN	46%	537	462	999
VEN	46%	14	12	26
MAR	42%	150	109	259
KOR	41%	78	55	133
TUR	41%	146	102	248
SGP	41%	23	16	39
MNE	40%	3	2	5
ZAF	39%	49	31	80
LUX	38%	178	109	287
RUS	38%	393	239	632
KWT	38%	10	6	16
BRA	37%	1394	830	2224
SVK	37%	41	24	65
ITA	35%	2433	1333	3766
BLR	35%	17	9	26
EGY	34%	21	11	32
HUN	33%	153	77	230
CIV	33%	4	2	6
ETH	33%	2	1	3
KEN	33%	4	2	6
LIE	33%	2	1	3
TMP	33%	2	1	3
COL	32%	48	23	71
THA	31%	41	18	59
NOR	30%	426	181	607
ECU	30%	19	8	27
UKR	29%	48	20	68
LBN	29%	22	9	31
MOZ	28%	48	19	67
URY	28%	23	9	32
MLT	28%	13	5	18
IRN	28%	60	23	83
TWN	27%	37	14	51
GRC	27%	93	35	128
ROU	27%	366	134	500
KAZ	26%	14	5	19
SVN	26%	42	15	57
ESP	25%	6391	2177	8568
ISR	25%	500	169	669
ARG	25%	160	54	214
AUS	25%	319	107	426
DNK	25%	326	109	435

## 7013SSL – Applied Marketing Analytics

ARM	25%	6	2	8
HRV	25%	75	25	100
MCO	25%	3	1	4
VNM	25%	6	2	8
CHE	25%	1302	428	1730
IRL	25%	2543	832	3375
CHL	25%	49	16	65
USA	24%	1596	501	2097
POL	23%	704	215	919
BIH	23%	10	3	13
IND	23%	117	35	152
OMN	22%	14	4	18
SWE	22%	797	227	1024
EST	22%	65	18	83
CZE	22%	134	37	171
CYP	22%	40	11	51
PER	21%	23	6	29
DZA	20%	82	21	103
BEL	20%	1868	474	2342
GBR	20%	9676	2453	12129
MKD	20%	8	2	10
CN	20%	1025	254	1279
FRA	19%	8481	1934	10415
NLD	18%	1717	387	2104
AUT	18%	1033	230	1263
DEU	17%	6069	1218	7287
ALB	17%	10	2	12
PRI	17%	10	2	12
LVA	16%	46	9	55
BGR	16%	63	12	75
FIN	15%	378	69	447
JOR	14%	18	3	21
MUS	14%	6	1	7
JPN	14%	169	28	197
NULL	14%	421	67	488
MEX	12%	75	10	85
GNB	11%	8	1	9
MYS	11%	25	3	28
LTU	9%	74	7	81
NZL	8%	68	6	74
ISL	7%	53	4	57
CRI	5%	18	1	19
SRB	3%	98	3	101
ABW	0%	2	0	2
AIA	0%	1	0	1
ASM	0%	1	0	1
ATA	0%	2	0	2
ATF	0%	1	0	1
BDI	0%	1	0	1
BFA	0%	1	0	1
BHS	0%	1	0	1
BOL	0%	10	0	10
BRB	0%	4	0	4
BWA	0%	1	0	1
CAF	0%	5	0	5
CMR	0%	10	0	10
COM	0%	2	0	2
CUB	0%	8	0	8
CYM	0%	1	0	1
DJI	0%	1	0	1
DMA	0%	1	0	1
GTM	0%	4	0	4



## 7013SSL – Applied Marketing Analytics

GUY	0%	1	0	1
IRQ	0%	14	0	14
JAM	0%	6	0	6
KIR	0%	1	0	1
KNA	0%	2	0	2
LAO	0%	2	0	2
LBY	0%	8	0	8
LCA	0%	1	0	1
LKA	0%	7	0	7
MDG	0%	1	0	1
MLI	0%	1	0	1
MMR	0%	1	0	1
MRT	0%	1	0	1
MWI	0%	2	0	2
NAM	0%	1	0	1
NCL	0%	1	0	1
NPL	0%	1	0	1
PAN	0%	9	0	9
PLW	0%	1	0	1
PRY	0%	4	0	4
PYF	0%	1	0	1
RWA	0%	2	0	2
SDN	0%	1	0	1
SLE	0%	1	0	1
SLV	0%	2	0	2
SMR	0%	1	0	1
STP	0%	2	0	2
SUR	0%	5	0	5
SYR	0%	3	0	3
TGO	0%	2	0	2
UGA	0%	2	0	2
<b>Grand Total</b>		<b>75166</b>	<b>44224</b>	<b>119390</b>

### Appendix 3.a

	B	C	D	E	F	G	H	I	J	K	L	M
1	is_cancelled	lead_time (X1)	adr (X2)	Forecast	Exponential Linear regression	P(x)	1-P(x)	Likelihood	Logarithm of likelihood (Ln)	Sum	Intercept	-2.367595
2	0	7	75	-1.608088	0.20027	0.166854	0.833146	0.833146	-0.18	-605.27	Slope 1	0.005798
3	0	13	75	-1.573302	0.207359	0.171746	0.828254	0.828254	-0.19		Slope 2	0.009586
4	0	14	98	-1.347034	0.26001	0.206356	0.793644	0.793644	-0.23			
5	0	14	98	-1.347034	0.26001	0.206356	0.793644	0.793644	-0.23			
6	0	0	107	-1.34193	0.261341	0.207193	0.792807	0.792807	-0.23			
7	0	9	103	-1.328094	0.264982	0.209475	0.790525	0.790525	-0.24			
8	1	85	82	-1.088772	0.33663	0.25185	0.74815	0.25185	-1.38			
9	1	75	105.5	-0.921486	0.397927	0.284655	0.715345	0.284655	-1.26			
10	1	23	123	-1.055214	0.348118	0.258225	0.741775	0.258225	-1.35			
11	0	35	145	-0.774758	0.460815	0.315451	0.684549	0.684549	-0.38			
12	0	68	97	-1.043547	0.352203	0.260466	0.739534	0.739534	-0.30			
13	0	18	154.77	-0.779666	0.458559	0.314392	0.685608	0.685608	-0.38			
14	0	37	94.71	-1.245225	0.287876	0.223528	0.776472	0.776472	-0.25			
15	0	68	97	-1.043547	0.352203	0.260466	0.739534	0.739534	-0.30			
16	0	37	97.5	-1.218481	0.295679	0.228204	0.771796	0.771796	-0.26			
17	0	12	88.2	-1.452569	0.233968	0.189607	0.810393	0.810393	-0.21			
18	0	0	107.42	-1.337904	0.262395	0.207855	0.792145	0.792145	-0.23			
19	0	7	153	-0.860407	0.42299	0.297254	0.702746	0.702746	-0.35			
20	0	37	97.29	-1.220494	0.295084	0.227849	0.772151	0.772151	-0.26			
21	0	72	84.67	-1.138548	0.320284	0.242587	0.757413	0.757413	-0.28			
22	0	72	84.67	-1.138548	0.320284	0.242587	0.757413	0.757413	-0.28			
23	0	72	99.67	-0.994763	0.369811	0.269972	0.730028	0.730028	-0.31			
24	0	127	94.95	-0.721137	0.486199	0.327143	0.672857	0.672857	-0.40			
25	0	78	63.6	-1.305732	0.270974	0.213202	0.786798	0.786798	-0.24			
26	0	48	79.5	-1.327249	0.265206	0.209615	0.790385	0.790385	-0.24			
27	1	60	107	-0.994072	0.370067	0.270109	0.729891	0.270109	-1.31			
28	0	77	94	-1.020125	0.36055	0.265003	0.734997	0.734997	-0.31			
29	0	99	87.3	-0.956801	0.38412	0.277519	0.722481	0.722481	-0.33			