Hotel Overbooking Strategy Using Cancellation Data

Joseph Despres
Rishabh Sareen

October 4, 2021

Abstract Overbooking is profitable in businesses such as airlines, cruise ships and even manufacturing. This project develops a profitable overbooking strategy using publicly available hotel cancellation data. The objective is to optimize the number of reservations given the uncertainty of a guest's arrival subject to an overbooking penalty. The optimal amount of bookings finds expected profit of the additional reservation no longer justifies the expected value of the overbooking penalty. This project shows, the number of empty rooms can be reduced by half, if a hotel is willing to pay modest overbooking penalties.

Work done in partial fulfillment of the requirements of Michigan State University MTH 843; advised by Dr. Peiru Wu, Michigan State University.

Table of Contents

Nomenclature

p	Probability of a guest arriving to a booking
A	Quantity of guests arriving to a reservation
b	Number of bookings made by the hotel
π	Profit
L	Loss function of penalty paid when arrivals exceed capacity
E	Expectation operator
$\hat{p}$	Estimated probability

Introduction

Overselling or overbooking is the practice of committing to sell more of a product or service than ability to fulfill. Airlines, oversell tickets and offer compensation in the event excess passengers arrive (14 CFR 250.2B). Manufactures often use similar algorithms to predict deferred or cancelled orders (Leahy 1). This permits them to accept more orders. However, this is a delicate task due to high costs incurred in the event of overselling.

If the probability of a guest arriving to a hotel is not 1, then the probability of a hotel operating with empty rooms is almost surely 1. This is where the opportunity to overbook presents itself. Implemented properly, overbooking offers a hotel the opportunity to increase revenue without incurring the high costs of increasing capacity. A profitable hotel overbooking strategy can be modified, repurposed, and increase revenues in a wide variety of industries.

The goal of this project is to construct a profitable overbooking strategy. Publicly available hotel booking data allows for validating this overbooking strategy and reporting performance. This strategy is constructed by making some assumptions about overbooking penalties, setting an arbitrary capacity, and estimating the uncertainty of each arrival. Then this strategy is tested with a simulation that resamples bookings until a predetermined capacity is reached.

The Data Section contains a description of the hotel cancellation data used. The Analysis Section explains the underlying mathematics The Methods Section contains the process of developing the overbooking strategy and the validation simulation. The Results section contains a discussion the results and a discussion of the applications and limitations. The Conclusion contains a summary of the results.

Data

This overbooking strategy is validated using a hotel booking demand data set (Antonio 1) in the public domain. These data are all hotel bookings from a hotel in Lisbon from 1st of July 2015 to 31st of august 2017. 119,930 observations detailing the hotel booking and cancellation records. The variables of interest are how many adults, children, and babies are included in a reservation, if a deposit was made, the days until reservation, the length of stay, and whether a deposit was made. If the booking is in a group or a in individual reservation. These variables were selected because in addition to being statistically significant predictors of cancellations, it is the information a booking agent would have access to at the time of a booking.

Basic descriptive statistics for the numeric variables can be found in Table 1. Find that on average 63% of bookings arrive. Children are rarely staying at this hotel and babies are rarely included in reservations. The average lead time is 104 days however the median is only 69 days indicating that the number of days customers book in advanced is right skewed. The high standard deviation suggests that there is significant variation in the lead time. Average stay length is approximately three and a half days, with a slight right skew and high standard deviation.

	Is canceled	Not canceled	Adults	Children	Babies	Lead time	Stay length
Minimum	0	0	0	0	0	0	0
Mean	0.37	0.63	1.86	0.1	0.01	104.01	3.43
Median	0	1	2	0	0	69	3
Standard Deviation	0.48	0.48	0.58	0.4	0.1	106.86	2.56

Table 1 Descriptive statistics on hotel arrivals and cancellations.

The categorical variables used to predict cancellations are the type of consumer and if a deposit was collected. The variable counts can be found in Table 2. Transient is a single guest unaffiliated with an organized group. This is the most common occurrence. Transient party is where an organized group has multiple bookings. Contract is when a booking has a formal business relationship with the hotel. Deposits are asked infrequently, and close inspection shows that deposits are taken in the event a customer is requesting flexible booking dates.

Customer type	Deposit	N
Transient	No Deposit	76,684
Transient-Party	No Deposit	23,858
Transient	Deposit	12,929
Contract	No Deposit	3,530
Transient-Party	Deposit	1,266
Group	No Deposit	569
Contract	Deposit	546

Table 2 Categorical variable counts

This dataset is used to validate this overbooking strategy. This project proceeds on the assumption the booking data are broadly representative, and there is no selection mechanism biasing the underlying dataset. Therefore, this project will proceed on the assumption that the hotel is not employing any overbooking strategy or selecting customers for the likelihood of arrival. Additional assumptions are that this data is representative of the arrival rates in the hotel industry.

Analysis

A guest making a hotel reservation is not certain to arrive. This uncertainty is defined as the probability of arrival p. A hotel will make many reservations and ask how many arrivals A to expect given number of bookings b?

The probability of b or fewer guests arriving follows a binomial distribution has a cumulative density function

$$\mathrm{F}(A|b,p) = \mathrm{P}(A \leq \mathrm{b}) = \sum_{\mathrm{i}=1}^{b} \binom{b}{i} \mathrm{p}^{\mathrm{i}}(1-\mathrm{p})^{b-\mathrm{i}}, \quad A \leq b, \quad 0 \leq p \leq 1.$$

This function gives the probability of A or fewer guests arriving given the number of bookings, and probability of arrival. With this distribution, the total profit a hotel can expect while booking no more than capacity is

$$E_b[\text{Total Profit}] = \pi b F(A|b,p),\ A \le b,\ \ 0 \le p \le 1,\ \ \pi > 0.$$

Where π denotes profit per room. Multiply profit, number of bookings, and sum probability of each arrival to get expected profit.

Booking over capacity allows fewer empty rooms, at the cost of risking insufficient rooms. Therefore, the costs of having more guests arrive than rooms available should be carefully considered. This strategy assumes that the cost of overbooking is providing each overbook additional accommodations at a cost of four times profit 4π. Now relax the assumption that bookings b is less than or equal to capacity k and obtain

$$E_b[\text{Total Profit}] = \pi b F(A|b,p) - 4\pi(b-k)\big(1 - F(A|b,p)\big),\ 0 \le p \le 1,\ \pi > 0.$$

This model estimates profit for a given number of bookings, accounting for the possibility of arrivals exceeding capacity. To determine the maximum number of bookings by finding the amount of booing's where $\boldsymbol{E_b}[\text{Total Profit}] > \boldsymbol{E_{b+1}}[\text{Total Profit}]$. Determine the number of bookings that maximize profit by solving this equation for b. Use an algorithm iterating over integer values of b starting at capacity and stopping when

$$\boldsymbol{E_b}[\text{Total Profit}] > \boldsymbol{E_{b+1}}[\text{Total Profit}].$$

See the source code in Table B.1 in Appendix for the implementation in R.

In the hotel booking data (Antonio 1), the average arrival rate of a reservation is 64%. Assuming a hotel has a 100-room capacity, the cost of overbooking is four times profit, and every guest has the same probability of arrival, maximize profit by booking 129 guests. The binomial distribution assumes that each p is the same for every trial. Assuming all customers have the same p, exposes the hotel to the inevitable event of booking too many customers highly likely to arrive and booking too few customers unlikely to arrive. The profitability of this strategy will depend on the ability to accurately estimate p. Therefore, this strategy requires an estimate $\hat{p}$.

The regression results in Table A.1 in Appendix A supports the claim each guest has a different probability of arrival. To estimate the probability each guest's arrival, fit a linear regression model to the log odds of arrival.

$$\log\left[\frac{\hat{p}}{1-\hat{p}}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots\ + \beta_n x_n.$$

Exponentiate both sides and solve for $\boldsymbol{\hat{p}}$

$$\hat{p} = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots\ + \beta_n x_n)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots\ + \beta_n x_n)}.$$

In addition to estimating $\boldsymbol{\hat{p}}$, the values of β_i estimate the effect of variable x_i.

Methods

It has been shown that different guests have different probabilities of arrival. This overbooking strategy begins by estimating the probability a given guest will arrive. The closer $\hat{p}$ is to the true value of p, the more profitable this strategy will be. Split the booking data into a training dataset and a testing data set. Rather than randomly selecting data, split the data by the first 80% chronologically. Then fit and validate a predictive model. A logistic regression model estimates the probability arrival to be

$$\log\left[\frac{\hat{p}}{1-\hat{p}}\right] = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\textbf{customer type}) + \boldsymbol{\beta_4}(\textbf{adults}) + \boldsymbol{\beta_5}(\textbf{children}) + \boldsymbol{\beta_6}(\textbf{babies}) + \boldsymbol{\beta_7}(\textbf{lead time}) + \boldsymbol{\beta_8}(\textbf{stay length}) + \boldsymbol{\beta_9}(\textbf{deposit}).$$

The estimated coefficients β_i can be found in Appendix A Table A.1. These regression coefficients are all statistically significant. This supports the claim that the probability of arrival $\hat{p}$ is not the same for each guest. Groups on average tend to have a higher arrival rate than individual short-term guests. The more adults in a single booking, the lower the expected arrival rate. Children seem to increase the odds of arrival until accounting for factors such as group, lead time, and stay length. Guests not making a deposit have a far higher arrival rate than those making non-refundable deposits. Closer inspection shows that non-refundable deposits are only asked of guests requesting flexible booking or have a history of cancelling reservations.

Use $\hat{p}$ to obtain the estimated fraction of capacity of the booking. Let b^* be the optimal number of bookings b where $E_{b+1}[\text{Total Profit}] - E_b[\text{Total Profit}] = 0$ given $\hat{p}$. Figure 1 shows that b^* goes to infinity as the probability of arrival approaches 0. In a commercial setting, this should be restricted to avoid the possibility of a catastrophic loss. This strategy assumes restricts bookings to no more than 10 times capacity.

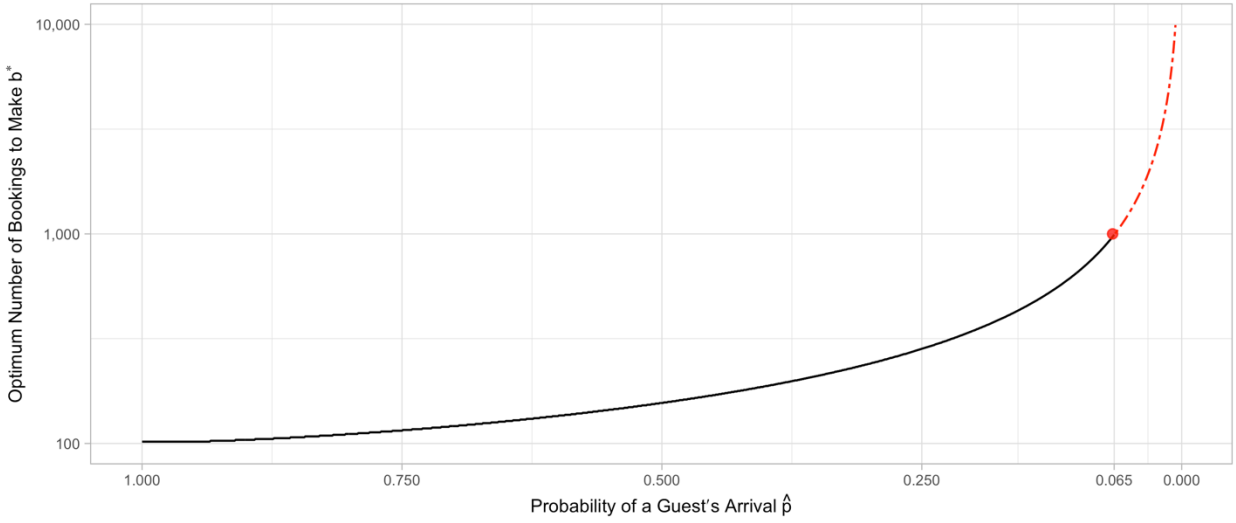

Figure 1 Relationship between estimated probability of arrival and optimum bookings.

A find the expected fraction of capacity C_f of a specific booking by taking the inverse of b^*

$$C_f = \frac{1}{b^*}.$$

Assign each booking an expected fraction of capacity. Then accept bookings until the total fraction of capacity exceeds 0.99.

$$\sum_i C_{fi} > 0.99$$

Randomly sample bookings with replacement on each date until estimated capacity is reached. For the details of the sampling function see the source code in Appendix A Table A.3.

To validate this method, a simulation is conducted to provide performance metrics. After estimating $\hat{p}$ on a testing set of data and determine b^*. Randomly sample guests with replacement until the hotel capacity is reached. This simulation assumes a hotel has 100 rooms. Then to study the results of $\hat{p}$ being the same sample as if $\hat{p} = 0.64$ and book 129 guests. This involves sampling with replacement until there are 129 guests in a day. Then sample on the b^* associated with each booking and sample for each day until $\sum_i C_{fi} > 0.99$. This results in an overbooking simulation of 162 days, where each day has bookings at capacity, overbooked assuming p is constant at 64%, and overbooking based on the estimate $\hat{p}$. Then is experiment is repeated 2,500 times.

Results and Discussion

Conclusions

References

"14 CFR 250.2B -- Carriers to Request Volunteers for Denied Boarding." *14 CFR 250.2b -- Carriers to Request Volunteers for Denied Boarding.*, https://www.ecfr.gov/current/title-14/chapter-II/subchapter-A/part-250/section-250.2b.

Antonio, Nuno, et al. "Hotel Booking Demand Datasets." *Data in Brief*, Elsevier, 29 Nov. 2018, https://www.sciencedirect.com/science/article/pii/S2352340918315191#f0010.

Leahy, John. "Too Many Orders? Yes, Says Consultant. No, Says Ex-Super-Salesman." *Leeham News and Analysis*, 28 Jan. 2019, https://leehamnews.com/2019/01/28/too-many-orders-yes-says-consultant-no-says-ex-super-salesman/.

Appendix A

The regression results in Table A.1 show statistically significant relationships between the log odds of a guest arriving and the covariates.

	Dependent variable:		
	Log Odds of Arrival		
	(1)	(2)	(3)
Adults	-0.183***	-0.065***	-0.231***
	(0.020)	(0.018)	(0.017)
Children	-0.161***	0.001	0.145***
	(0.025)	(0.025)	(0.024)
Babies	0.730***	0.890***	1.236***
	(0.154)	(0.156)	(0.153)
Lead Time (Days)	-0.004***	-0.007***	
	(0.0001)	(0.0001)	
Stay Length (Days)	-0.076***	-0.030***	
	(0.005)	(0.005)	
Customer Type: Group	1.312***		
	(0.220)		
Customer Type: Transient	0.061		
	(0.054)		
Customer Type: Tranient-Party	0.582***		
	(0.057)		
No Deposit	6.295***		
	(0.187)		
Constant	-4.654***	1.313***	0.764***
	(0.197)	(0.036)	(0.032)
Observations	59,494	59,494	59,494
Log Likelihood	-28,857.340	-36,311.220	-40,138.290
Akaike Inf. Crit.	57,734.680	72,634.440	80,284.570
Note:			*p<0.1; **p<0.05; ***p<0.01

Table A.1 Regression results showing that the probability of arrival is unique to each guest.

Appendix B

Table B.1 contains the source code for implementing the overbooking strategy. Input phat capacity, loss function, and the minimum probability this is to calculate. After initializing some values, this function takes a given p and estimates the expected units of profit given that probability of arrival while accounting for the possibility of too many guests arriving.

```
capacity_frac ← function (p, rooms=100, loss=4, min_p=0.065) {

  # input phat associated with each booking
  E ← 0; E_m1 ← 0; Fx ← 0; P_overbook ← 0; bookings ← rooms;
  # Initialize values

  if(p > min_p) { # set a maximum overbooking threshold
    while(E_m1 ⩾ E) {
      E ←  bookings * Fx - loss * ((bookings - rooms) * P_overbook)
      Fx ←  pbinom(rooms, bookings, p, lower.tail = TRUE)
      bookings ←  bookings + 1
      P_overbook ←  1 - Fx
      E_m1 ←  bookings * Fx - loss * ((bookings - rooms) * P_overbook)
      }
  } else {
    bookings = rooms * 10
  }
  return(bookings)
}

test_pred ← test %>%
  mutate(p_hat = predict(logistic_fit,
                         newdata = test, type = "response"),
         cap = map_dbl(p_hat, capacity_frac),
         cap_fra = 1/cap)
```

Table B.1 Source code for the function converting estimated probability of arrival to fraction of capacity.

Table B.2 Contains the resampling functions for the three scenarios that are tested. First, a unique estimate for *p*. The input is a single day of a hotel and samples from that until the total expected fraction of capacity exceeds 0.99. In other words, this simulates the process of booking and overbooking hotel reservations until the capacity is reached weighting the guests based on their probability of arrival.

```r
sample_fn_overbook ← function(data_input, p_threash=0.99) {
  # Input data for a single booking date
  success ← FALSE
  i ← 1
  samp ← data_input[1,]
  while(!success) {
    samp[i,] ← slice_sample(data_input, n = 1) # randomly select booking until
    success ← sum(samp$cap_fra) > p_threash    # expected capacity is reached
    i ← i + 1
  }
  samp ← samp[1:i,]
  # output resampled bookings
  return(samp)
}

sample_capacity ← function(data_input, capacity = 100) {
  # Input data for a single booking date
  output ← slice_sample(data_input, n = capacity, replace = TRUE)
  # Sample to hotel capacity with replacement
  return(output)
}

sample_same_phat ← function(data_input, capacity = 100) {
  # Input data for a single booking date
  output ← slice_sample(data_input, n = 129, replace = TRUE)
  # Sample 129 rooms assuming each guest has the same
  # likelihood of arrival
  return(output)
}
```

Table B.2 Resampling functions that simulate the overbooking strategy.

Table B.3 Contains the function that implements multiple iterations of this resampling process. This runs the resampling process for each of the testing days 2,500 times.

```r
simulater <- function (x) {
  message(paste0(x))
  test_pred %>%
    group_by(check_in_date) %>%
    nest() %>%
    mutate(
      resampled_overbookings = map(data, sample_fn_overbook),
      resampled_same_phat = map(data, sample_same_phat),
      resampled_capacity = map(data, sample_capacity),

      over_nbookings = map_dbl(resampled_overbookings, ~nrow(.x)),
      over_canceled = map_dbl(resampled_overbookings, ~sum(.x$is_canceled, na.rm = TRUE)),
      over_booking_arrivals = over_nbookings - over_canceled,

      over_same_phat = 129,
      over_canceled_same_p = map_dbl(resampled_same_phat, ~sum(.x$is_canceled, na.rm = TRUE)),
      over_booking_arrivals_phat = over_nbookings - over_canceled,

      booking_n = 100,
      booking_canceled = map_dbl(resampled_capacity, ~sum(.x$is_canceled, na.rm = TRUE)),
      booking_arrivals = booking_n - booking_canceled,

      iter = x
    )   %>%
    select_if(is.numeric)
}

map(1:2.5e3, simulater) %>%
  # run simulation 2500 times
  # note this requires approximately
  # 18 hours of computer time
  bind_rows() %>%
  select_if(is.numeric) %>%
  write.csv(here::here("data", "sim_results_iterations.csv"))
```

Table B.3 Simulation function that repeats the resampling functions for 2500 reps.

If you wish to recreate these results find the full project source code here.