A Statistical Approach to Hotel Overbooking

Joseph Despres
Rishabh Sareen

September 23, 2021

Abstract Overbooking has a wide range of commercial applications from airlines to concert halls. This project outlines a strategy for hotels to maximize bookings based on commercial hotel data. The task is to develop a profitable overbooking strategy, given the uncertainty of a reservation's arrival and the cost of providing alternative accommodations. First, a logistic regression model estimates the probability of cancellation given days until check in, booking agency, deposit, price paid, etc. Then a binomial probability model estimates the maximum number of reservations to schedule given the number of rooms and cost of overbooking. Splitting data into training and testing sets, shows strategy is TBD more profitable than abstaining from overbooking. [Joe wrote, Rishabh suggested edits]

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
E	Expected value
β	Number of bookings made
η	Profit per room
F	Cumulative density function (CDF) of a binomial random variable
ℓ	Cost of purchasing alternative accommodations

Introduction

Many businesses such as airlines, cruise ships, concert halls and hotels benefit from some type of overbooking. Customers purchasing these services tend to have a non-zero probability of canceling in advanced or missing the reservation. Regardless of whether a business takes payment in advanced, collects a deposit, or schedules without a no-show penalty, overbooking is worthy of consideration. However, there are high costs associated with selling more than can be provided such as alternative accommodations, reputational damage, irate customers, and so on. Therefore, overbooking must be done with care and strong mathematical justification. Probability models give us the proper tools to design a profitable overbooking strategy.

The optimal overbooking strategy schedules enough reservations so that the business is near or at capacity, but rarely incurs the costs of overscheduling. The optimum number of people is based on the probability of them arriving and the costs associated scheduling more than can be fulfilled. Non-monetary considerations like reputation or regulations are important however, this strategy will only consider cost. This model is based on real hotel booking dataset which can later be generalized and adjusted to another business's requirements.

The starting point is to train a classification model that predicts the probability of an individual guest arriving because each guest has a unique probability of arrival. This will use relevant factors such time until booking, number of guests in the reservation, agency, and month of booking to estimate the probability of arrival. Once the individual probability is estimated it is inputted into a binomial model that estimates the number of bookings considering the increase in profit from an additional booking is no longer justified given the additional risk.

This project begins with a discussion of the data sources used to build this model. Then in the Analysis section we justify the use of logistic regression in conjunction with a binomial model. The methods section contains an introduction of the functions that combine these models on one data source and includes a simulation of random overbooking to benchmark our overbooking strategy against doing no overbooking or taking every booking. After that a discussion of the results, implications and how this process can be generalized to any business taking sufficiently large reservations and how many reservations must be taken to justify an overbooking strategy. This project concludes with a summary of the results, limitations, and other considerations.

[Joe and Rishabh wrote together]

Data

The data used in this project is provided directly from two anonymous hotels in Portugal. As commercial data is often difficult to obtain, this dataset was specifically released to assist in the development of models that can more accurately predict cancellations. Real data gives us the ability to estimate probabilities and test the performance. These data are a combination of two datasets one from a resort hotel and another from a city hotel. Combined they contain roughly 120,000 observations. Each observation represents a hotel booking and most importantly informs us if the reservation was canceled. Although there are 32 variables in the dataset published, the model selection process only uses eight of them, which are explained in Table 1.
[Joe made the table and Rishabh wrote the narrative then edited jointly]

Table 1 Hotel booking data description

Variable	Type	Description
is_canceled	Categorical	Indicates if the reservation cancels their booking.
lead_time	Numeric	The number of days until the reservation.
arrival_month	Categorial	The month of the booking.
stay_length	Numeric	Length of stay the guest schedules.
is_repeated_guest	Categorical	Informs if the guest has stayed at the hotel in the past.
previous_cancellations	Numeric	Number of times guest has cancelled reservations.
deposit_type	Categorical	Indicates if a customer made a deposit.
adults	Numeric	Number of Adults on the reservation.

Analysis

There are many good options for classifying binary outcomes given a dataset. The advantage of logistic regression is interpretability and strong theoretical justifications. This is an adaptation of ordinary least square (OLS). Under a few assumptions, it can be shown that OLS estimators are the best linear unbiased estimators. OLS estimators obtain a slope that minimizes the sum of the squared distance between each point and its predicted value. However, OLS has a prediction range of $(-\infty, \infty)$ and since probability is strictly defined on [0, 1], this must be adapted. To avoid a function outputting an undefined probability estimate, convert probability to odds and take the logarithm. Odds is defined as $\frac{p}{1-p}$, with a domain of $[0, \infty)$ the logarithm of this ratio has a range on the real line. Use OLS on the odds ratio to estimate the probability of being in each category.

This method is widely used in classification problems. Although machine learning methods have proven more accurate, logistic regression has a stronger theoretical justification for the probability estimates, uncertainty quantification and estimated prediction intervals. Since this probability estimate is being used as an input into another function, theoretical justification and interpretability is preferred over out of sample accuracy.

A hotel guest arriving to a reservation follows a binomial distribution. The profit from each booking is modeled by multiplying the number of bookings by profit and the CDF of the binomial distribution for the number of bookings, rooms, and probability. If the number of bookings does not exceed rooms the following loss function will be equal to zero. Take the compliment of the CDF to obtain the probability of overbooking, multiply that by the number of bookings minus rooms times the cost of additional accommodations. This is expressed as

$$E_{\beta}(\eta) = \beta\eta F(\beta) - ((\beta - \delta)\eta\, \ell(1 - F(\beta))).$$

This is a function of profit per room, hotel capacity, probability of arrival, cost of overbooking and number of bookings made. Profit per room and cost of alternative accommodations are fixed constants. A strong case could be made that the cost of booking last minute alternative accommodations is not a linear function. In general, it seems that the cost of purchasing additional accommodation can be modeled as a positive multiple of the profit per room.

After establishing the function, the aim is to find the number of bookings where $E_{\beta+1}(\eta) - E_{\beta}(\eta) = 0$ for a fixed profit, capacity, and cost of additional accommodations. This can be done using iterative methods. Start by initializing the constant values and iterate until this equation equals zero. This maximizes bookings given a probability of arrival. This generates a vector of the estimated maximum number of bookings to take given the individual's probability of arrival. Since each guest has a different probability of arrival, the maximum number of bookings will be different in each case. This is a vector of which the inverse is the fraction of the hotel capacity the guest is expected to take. Now, each booking has an estimated fraction of capacity. When a booking is made add the fraction of capacity until the capacity is reached.

This overbooking strategy will be tested on the hotel booking dataset. A simulation of this method is performed on a testing partition. Since marketing is outside the scope of this project, we simulate the scenario where guests continuously requesting reservations. If the testing set does not contain enough bookings bootstrap resampling method will generate additional guests. This will give an idea of the type of performance to expect, how much more profit an overbooking strategy can yield and the minimum capacity threshold where an overbooking strategy is useful.

[Joe and Rishabh discussed this together, Joe wrote the logistic regression part and MATLAB algorithm, Rishabh suggested edits]

Methods (or Procedure)

Results and Discussion

Conclusions

- Each guest has a unique, but predictable probability of arrival, therefore any overbooking strategy should account for that.
- Using a simulation, employing this hotel overbooking strategy is TBD more profitable than not employing an overbooking strategy.
- An overbooking strategy is commercially viable only when the number of rooms is greater than TBD.

[Joe and Rishabh discussed this together, Joe wrote the logistic regression part and MATLAB algorithm, Rishabh suggested edits]

References [This section must start on a new page. Please stay with uniform format (e.g., MLA format) for the references list below.]

Appendix A [Each appendix section must start on a new page.]

Appendix B [Each appendix section must start on a new page in case you have more than one appendix section.]