**A Commercially Viability, Data Driven Hotel Overbooking Strategy**

Joseph Despres

Rishabh Sareen

September 16, 2021

**Abstract** Commercial hotel booking data guides a probability model designed to optimize hotel overbooking. The task is to develop a profitable overbooking strategy, given the uncertainty of a reservation’s arrival and the cost of providing alternative accommodations. Then a logistic regression model estimates the probability of cancellation given days until check in, booking agency, deposit, price paid, etc. The probability model in conjunction with logistic regression shows the maximum number of additional reservations to schedule. Splitting data into training and testing sets, shows this over booking strategy is TBD more profitable than accepting every booking, and TBD more profitable than abstaining from overbooking.

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** [This is an example (including pages numbers) to give you some ideas about a report structure.]

[Nomenclature 1](#_Toc36765862)

[Introduction 2](#_Toc36765863)

[Data](#_Toc36765864) 3

[Analysis 4](#_Toc36765864)

[Methods (Procedure)](#_Toc36765866) 6

[Results and Discussion 8](#_Toc36765867)

[Conclusions 12](#_Toc36765868)

[References 14](#_Toc36765871)

[Appendix A 15](#_Toc36765872)

[Appendix B 18](#_Toc36765873)

# Nomenclature (*In order of its appearance*)

# 

# 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, collected a deposit, or schedules without a no-show penalty, overbooking is worthy of consideration. However, there are high costs associated with selling too many spaces such as providing 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 faces costs of overscheduling. The optimum number of people is based on the probability of them arriving and the costs paid to those that will be turned away. This will be different for every business and non-monetary considerations are important this model will only consider cost.

Although a profitable strategy could be derived purely theoretically, this model will be built surrounding hotel booking data. That can be generalized to find various businesses. The starting point is to train a logistic regression model and predict the probability of a given guest arriving to the hotel for their booking. This will use relevant factors such time until booking, guests, agency, and blah. To predict the probability of a guest arriving. More powerful classification engines are of course available, but MATLAB sucks and that would be a pain to code, but this part can be exchanged without trouble. The induvial probability is inputted into a binomials model that estimates the number of bookings where the increase in profit no longer justifies another booking

This project beings with a discussion of the data sources we use 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, limititations, and other considerations.

# Data

# Due to its value, real commercial data is difficult to obtain. The dataset is 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. These datasets were specifically released to assist in the development of models that can more accurately predict cancellations. Although there are 32 variables in the dataset published, the model selection process only uses 8 of them. Which are explained in Table 1. This is plenty sufficient to construct an overbooking strategy.

|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| is\_canceled | Categorical | Indicates if the reservation cancels their booking. |
| b |  |  |
|  |  |  |

# Analysis

# Methods (or Procedure)

# Results and Discussion

# Conclusions

# 

# 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.]