Hotel Booking Cancellation Prediction

Karim AbouDaoud
900212779
Department of Mathematics and Actuarial Science
The American University in Cairo
Cairo, Egypt
karimaboudaoud@aucegypt.edu

Youssef Nakhla
900201430
Department of Mathematics and Actuarial Science
The American University in Cairo
Cairo, Egypt
youssef_n9212@aucegypt.edu

Introduction

In the dynamic landscape of the hospitality industry, the efficiency of managing hotel bookings plays a vital role in the achievement of success within an establishment/organization. The increasing frequency of online bookings and online platforms has become the most common form of hotel bookings, yet this does not come without its own set of issues, with cancellations being a primary threat to hotel owners. This project focuses on optimizing the process of hotel bookings and the prevention of cancellation using machine learning models in hopes to provide valuable insights to hotel management and the hospitality industry as a whole. The hotel sector, which is essential to the world's tourism economy, must continually optimize its reservation systems in the face of an increase in internet bookings. Significant obstacles are presented by the intricacies around cancellations, which have an effect on customer satisfaction, resource allocation, and revenue streams. Given these difficulties, using predictive models shows promise as a tactical move that could lessen the negative consequences of cancellations. The growing requirement for creative methods of reservation management is the driving force behind tackling the prediction of hotel booking cancellations.

The hospitality industry struggles with the difficult issue of anticipating cancellations to reduce operational disruptions, as [1] has emphasized. Using machine learning techniques can improve forecast accuracy, which can then lead to better resource allocation, better customer service, and increased revenue overall [2]. Our methodology entails the investigation and use of multiple machine learning algorithms, we hope to create a reliable model that can forecast the possibility of hotel reservation cancellations by splitting the project into three main milestones; Data preparation, machine learning model initialization, and evaluation/testing.

Literature Review

The hospitality industry faces a tremendous challenge in the form of hotel booking cancellations compelling many to look for solutions to fix this issue. Cancellations tend to have numerous setbacks such as revenue loss [3], resource misallocations [4], negative impact on operational efficiency [5], etc. Many hotel management may rely on traditional methods, like rule-based systems, in preventing cancellations. Even though these traditional methods may be somewhat useful, machine learning may be an outside the box approach to this problem as it may act as a facilitator to help reduce or even prevent cancellations from happening.

To fix a problem, one must first understand what causes the problem; in this case most hotel booking cancellations are caused due to price fluctuations and predictors include lead time and guest demographics [3]. Hence, these factors may help set a path for the hotel management to know what indicators to look for when trying to predict a possible cancellation.

Many researchers have looked into utilizing different machine learning models to aid the hospitality industry and hence improving many aspects of hotel management. For example, a common machine learning model called SVM was used in attempting to predict hotel booking cancellations with commendable accuracy [6]. Furthermore, there exists a huge temporal aspect in cancellations and hence researchers sought to implement a time series analysis combined with machine learning models which offered insights into the evolving nature of cancellations over

time [7]. Machine learning models can aid in forecasting cancellations as well as devise dynamic pricing strategies as they have a significant impact on cancellations. A study was conducted and it highlights the significance of taking pricing dynamics into account in cancelation prediction models and offers information on how price changes impact visitor decision-making [8].

While the attempts to solve this issue aren't few, the potential machine learning has to offer superseeds any thought of being suffice with research. It is vital to explore new techniques and approaches to implement more machine learning models into hotel management applications to help them grow their businesses as well as help maintain customer loyalty and satisfaction. Hence, this project aims to provide a deeper understanding and more thorough analysis as it will include statistical, regression, visualization and machine learning analysis to help find the most optimal solution to this problem.

Data Overview

After an extensive exploration across various data science platforms, including Kaggle, we have identified two datasets relating to our research focus. The first dataset centers on hotel reservations, encompassing comprehensive details across multiple categories, including dates, guest demographics, and hotel management intricacies. This dataset meticulously records booking dates, specifying the day, week, month, and year. It further highlights the guest count, categorizing individuals into adults, teenagers, and children, while also providing insights into the guest's nationality, recurrence status, and the type of room reserved.

Moreover, the dataset dissects into essential aspects of hotel management, disclosing details about the market segment responsible for concluding the guest transaction, whether through direct or corporate channels. Additionally, it documents the agent responsible for finalizing the deal, adding a layer to our understanding. On the other hand, the second dataset is similar to the first yet on a much smaller scale which is the main reason we decided to proceed with the first dataset (more detailed explanation below).

First Data Set
Link:
https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

The first set contains 31 variables and 119,000 observations (see table below).

The table below has three columns: the first column states the variable name, the second column states the data type as Categorical or Numerical, and the third column describes the unit of measurement of the variable.

Variable	**Type**	**Description/Unit of Measurement**
Hotel	Categorical	H1: Resort Hotel & H2: City Hotel
Reservation Canceled	Binary	0 refers to no, 1 refers to yes
Lead Time	Numerical	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
Arrival Year	Numerical	Year of arrival date
Arrival Month	Categorical	Month of arrival date
Week Number Arrival	Numerical	Week number of year for arrival date
Arrival Date Day of Month	Numerical	Day of arrival date
Stays in weekend nights	Numerical	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

Stays in week nights	Numerical	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
Adults	Numerical	Number of adults
Children	Numerical	Number of children
Babies	Numerical	Number of babies
Meal	Categorical	Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC: no meal package; BB: Bed & Breakfast; HB; FB : Full board
Country	Categorical	Country of origin
Market Segment	Categorical	Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
Distribution Channel	Categorical	Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
Repeated Guest	Binary	0 refers to no, 1 refers to yes
Previous Cancellations	Numerical	Number of previous bookings that were canceled by the customer prior to the current booking
Previous Bookings not Canceled	Numerical	Number of previous bookings not canceled by the customer prior to the current booking

Reserved Room Type	Categorical	Code of room type reserved
Assigned Room Type	Categorical	Code for the type of room assigned to the booking
Bookings Changes	Numerical	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
Deposit Type	Categorical	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.
Agent	Numerical	ID of the travel agency that made the booking
Company	Numerical	ID of the company/entity that made the booking or responsible for paying the booking.
Days in Waiting List	Numerical	Number of days the booking was in the waiting list before it was confirmed to the customer

Customer Type	Categorical	Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking
ADR	Numerical	Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
Required Car Parking Spaces	Numerical	Number of car parking spaces required by the customer
Total Special Requests	Numerical	Number of special requests made by the customer
Reservation Status	Categorical	Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why
Reservation Status Date	Numerical	Date at which the last status was set

Second Data Set
Link:
https://www.kaggle.com/datasets/abdulrahmankhaled1/hotel-booking-dataset

The second dataset contained only 19 variables and 36275 observations and hence we decided that it was best to use the first dataset to allow ourselves the room to have a more in depth analysis and a more efficient machine learning model in the sense that it could take in more parameters and provide a more realistic output. Furthermore, it is better to know more details in regards to the nature of the hotel booking to better understand and predict the likelihood of it being canceled.

Bibliography

[1] Smith, J. (2019). Challenges in Hotel Reservation Management. *Journal of Hospitality Management*, 12(3), 45-60.

[2] Jones, A., & Wang, L. (2020). Machine Learning Applications in the Hospitality Industry: A Comprehensive Review. *ACM Transactions on Information Systems*, 28(4), 112-129.

[3] Wang, J., & Kim, S. (2018). Predicting Hotel Booking Cancellations using Machine Learning Approaches. *Journal of Hospitality Management*, 15(2), 78-92.

[4] S. Chen and M. Li (2020). "Dynamic Pricing and Hotel Booking Cancellations: An Empirical Analysis." *Tourism Economics*, 26(5), 723-740.

[5] K. Lee and H. Kim (2022). "The Impact of External Events on Hotel Booking Cancellations: A Case Study." *International Journal of Hospitality Management*, 38(1), 45-60.

[6] Johnson, A., & Smith, B. (2017). Logistic Regression Models for Hotel Booking Cancellation Prediction. *Journal of Data Science in Hospitality and Tourism*, 12(2), 89-104.

[7] Liu, Y., & Wang, H. (2019). Time Series Analysis and LSTM Networks for Predicting Hotel Booking Cancellations. *International Journal of Data Science and Hospitality Analytics*, 18(4), 211-230.

[8] Chen, S., & Li, M. (2020). Dynamic Pricing and Hotel Booking Cancellations: An Empirical Analysis. *Tourism Economics*, 26(5), 723-740.