### Team Members: Ashisha Konnur, Tao Ma, Tobias Sundheimer

### BAN 5600 Deliverable 1

### Due: 10/16/21

Hotel Booking Demand Dataset

Abstract

Have you ever wondered when is the best time of the year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? Accurate forecasting of the hotel accommodation demands is extremely critical to the sustainable development of tourism-related industries. In view of the ever-increasing tourism data, this paper constructs a deep learning framework to handle the prediction problem in the hotel accommodation demands. We are going to assess the Hotel Demand of the 2 cities in Portugal. Forecasting results can provide necessary references for decision-making in tourism-related industries, and this forecasting framework can also be extended to other similar complex time series forecasting problems. One of the hotels (H1) is a resort hotel and the other is a city hotel (H2). Both datasets share the same structure, with 31 variables describing the 40,060 observations of H1 and 79,330 observations of H2. Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled.

***Keywords***: internet search index; deep learning framework; LSTM model; hotel accommodation demands; forecasting performance

Introduction

The following dataset we will be working on is on Hotel Demand in Lisbon and Algarve, Portugal. Since this is hotel real data, all data elements pertaining hotel or costumer identification were deleted. Due to the scarcity of real business data for scientific and educational purposes, these datasets can have an important role for research and education in revenue management, machine learning, or data mining, as well as in other fields.

## Problem statement

In tourism and travel industries, most of the research on revenue management demand forecasting and predication problems employ data from the aviation industry to predict the hotel demand.

The dataset now made available is aimed at the development of

Descriptive Analytics:

* Where do the guests come from?
* How much do guests pay for a room per night?
* How does the price per night vary over the year?
* How Many Booking Were Cancelled?
* What is the booking ratio between Resort Hotel and City Hotel?
* What is the percentage of booking for each year?
* Which is the busiest month for hotels?
* From which country most guests come?
* How Long People Stay in the hotel?
* Which was the most booked accommodation type (Single, Couple, Family)?

Predictive Analytics:

* Have prediction models to classify a hotel bookings likelihood to be canceled.
* We may also want to predict price surge or contract with seasonality and variables in our dataset.

## Map Description automatically generatedData source: Hotel Bookings

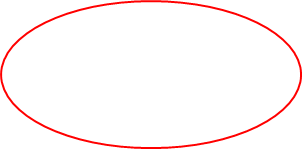
We sourced our data from Kaggle[[1]](#footnote-2). This dataset contains 119,000 observations from two hotels, one in Lisbon, and one in the resort region of Portugal (Algarve). The dataset comprehends bookings due to arrive between the 1st July of 2015 to 31st August of 2017, which will also give us information on the season ebbs and flows of the tourist industry. While there are missing values in this dataset, the missing values only constitute a negligible portion of the dataset itself and are randomly distributed, so we can drop those observations without skewing our information. The only column that is missing significant amounts of values in it is the “company” column which makes sense, since only corporate customers will have an affiliated company with them. We can disregard this column as it does not contain any useful information for us. Even after dropping this column, we have about 30 different qualitative and quantitative metrics available to us, which is more than enough information for us to generate a holistic predictive model on the likelihood that someone will cancel their hotel bookings at either location.

Figure 1 - Location of Lisbon and Algarve in Portugal.

## Data Dictionary:

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| **hotel** | Hotel (Resort Hotel or City Hotel) |
| **is\_canceled** | Value indicating if the booking was canceled (1) or not (0) |
| **lead\_time** | Number of days that elapsed between the entering date of the booking into the PMS and the arrival date |
| **arrival\_date\_year** | Year of arrival date |
| **arrival\_date\_month** | Month of arrival date |
| **arrival\_date\_week\_number** | Week number of year for arrival date |
| **arrival\_date\_day\_of\_month** | Day of arrival date |
| **stays\_in\_weekend\_nights** | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel |
| **stays\_in\_week\_nights** | Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel |
| **adults** | Number of adults |
| **children** | Number of children |
| **babies** | Number of babies |
| **meal** | Type of meal booked.  Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner) |
| **country** | Country of origin. (PRT-Portugal, GBR-UK, etc…) |
| **market\_segment** | Market segment designation.  (TA-Travel Agents, TO- Tour Operators) |
| **distribution\_channel** | Booking distribution channel.  (TA-Travel Agents, TO- Tour Operators) |
| **is\_repeated\_guest** | Value indicating if the booking name was from a repeated guest (1) or not (0) |
| **previous\_cancellations** | Number of previous bookings that were cancelled by the customer prior to the current booking |
| **previous\_bookings\_not\_canceled** | Number of previous bookings not cancelled by the customer prior to the current booking |
| **reserved\_room\_type** | Code of room type reserved. |
| **assigned\_room\_type** | Code for the type of room assigned to the booking. |
| **booking\_changes** | 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** | Indication on if the customer made a deposit to guarantee the booking. (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** | ID of the travel agency that made the booking |
| **company** | ID of the company/entity that made the booking or responsible for paying the booking. |
| **days\_in\_waiting\_list** | Number of days the booking was in the waiting list before it was confirmed to the customer |
| **customer\_type** | Type of booking (Group, Contract, Transient, Transient-Party) |
| **adr** | Average Daily Rate |
| **required\_car\_parking\_spaces** | Number of car parking spaces required by the customer |
| **total\_of\_special\_requests** | Number of special requests made by the customer |
| **reservation\_status** | Reservation last status (Canceled; Check-Out; No show) |
| **reservation\_status\_date** | Date at which the last status was set. |

Text

Description automatically generated

Figure 2 - data frame description

Literature Review

Trying to solve the problem on how to predict hotel room attrition is not a new one, and there are many theories on how to accurately predict how expensive a room will cost. If we are able to correctly predict which groups of people are regularly cancelling their bookings as well as why they are doing so, we can then design campaigns to entice them to follow through on their bookings. For a hotel, having reserved rooms going unused represents one of the largest losses of revenue possible, since that room COULD have been used for a paying customer, but instead is going empty because the person who originally claimed it ended up not showing up.

Multiple different resources all point to location as being an extremely important factor when determining hotel price[[2]](#footnote-3). Since both the areas that we are studying have multiple tourist attractions in their relative proximity but display different prices as well as different rates of booking cancellations, we can determine what types of attractions gets people to actually follow through on their bookings, as well as increase the lengths of their trips.

A study was conducted in 2018, published in the International Journal of Contemporary Hospitality Management, on the determinants of cancellations by hotel guests[[3]](#footnote-4). In their study, they found that the most important factors were the lead time (how long in advance a person books their stay vs when they are scheduled to stay at the hotel) and the country of origin. Lead time intuitively makes sense, since the longer amount of time between booking and the actual stay gives more time for life to get in the way and plans to need to shift. Something more interesting for us to try to look into is whether or not these people ever re-book at the same hotel, or if they are truly attritted customers. The differences across country of origins is also interesting, since the countries that have higher cancellation rates are the ones where you have to file much more paperwork to cross boarders, so people may be cancelling in higher rates because they fail to complete the applications needed to take their trip in the first place. The final main takeaway from this study is that the different types of trips (business vs leisure) have different cancellation rates, with business trips having lower cancellations than vacation trips. Even though we are looking at primarily tourist destinations, we will also be able to check if there are any statistically significant differences between business and leisure bookings.

The National Consumer’s League (NCL) conducted another study about the cancellation rates in American hotels, and found that the booking method also played a large role in determining whether or not someone was going to cancel their hotel stay[[4]](#footnote-5). They found that compared to direct bookings (via hotel specific websites), travel agency bookings had twice the number of cancellations over the given timeframe. This means that using Online Travel Agencies may not be the best idea to fill rooms, and that hotel owners should be focusing their resources on other outreach methods, and funneling more customers directly into their site itself.

visualization Analysis

**Overall Hotel Cancellation Rate Analysis**

The figure 3 shows that the overall number of bookings for city hotels is much higher than for resort hotels, almost twice as much. We speculate that there are more local city hotels than resort hotels, in addition to the fact that city hotels are cheaper and therefore in higher demand. In addition, the overall cancellation rate, regardless of hotel type, was 37.04%, so we further compared whether there was a significant difference in the cancellation rate between city and resort hotels.

Chart, pie chart

Description automatically generated Chart, pie chart

Description automatically generated

Figure 3 - Hotel Types and Cancellation Rates of the Dataset

From Figure 4, we can see that City hotels have a higher cancellation rate than resort hotels, which may be related to the different services, prices, and functions offered by the two hotels. In general, both these hotel types have a high cancellation rate.

Chart, bar chart

Description automatically generated

Figure 4 - City Hotels and Resorts Cancellation

Figure 5 shows the data from July 2015 to August 2017, from which we can see that April-June is a peak booking period and has the highest cancellation rate every year. In addition, the booking rate is also high from September to October and the cancellation rate is also high.

Chart, histogram

Description automatically generated

Figure 5 - Monthly average order amount and Cancellation rate

**Customer Profile Analysis**

*Customer Source Analysis*

Figure 6 shows that the majority of customers are local from Portugal, while French and German guests prefer city hotels, and UK and Spanish guests prefer resort hotels.

Chart

Description automatically generated Chart

Description automatically generated

Figure 6 - Resort and City Hotels Guest Countries

From the market segments in Figure 7 both types of hotels derive most orders from online travel agents, followed by offline TA/TO.

Chart, pie chart

Description automatically generated Chart, pie chart

Description automatically generated

Figure 7 - Resort and City Hotels Guest Market Segments

*User Demand Attributes*

From Figure 8, we observe that the probability of cancellation within 0-10 days is relatively small, and after 30 days, the cancellation probability gradually increases, hence hotels can consider shortening the advance booking period.

Chart, histogram

Description automatically generated

Figure 8 - Cancellation rate for the number of days between booking and check-in

* Meal Type

From Figure 9, we find that Full Board has the highest cancellation rate, thus we consider that Full Board may be driving up the overall order price and that there is not high demand from guests for Full Board. Therefore, the hotel could consider focusing on Breakfast and Half Board and provide some special offers.

Chart, line chart

Description automatically generated

Figure 9 - Meal types and Cancellation rates

* Room Type

Figure 10 shows that the demand for the A type room is the highest and the cancellation rate is low, while the demand for the P type room is almost zero and the cancellation rate is high. Therefore, we believe that the hotel can consider upgrading the A room type to provide better service to customers, and at the same time make changes to the P type room to increase the demand rate and reduce the cancellation rate.

Chart, line chart

Description automatically generated

Figure 10 - Room types and Cancellation rates

**Customer Behavior Analysis**

Figure 11 shows that there is no direct correlation between the number of order changes and final cancellations, but the relative cancellation rate is lower for those who have made changes, therefore the hotel may consider trying to implement an order change option.

Chart, line chart

Description automatically generated

Figure 11 – Change times and Cancellation rates

**Sales Analysis**

Figure 12 shows the hotel sales data from July 2015 to August 2017, and we observe that the city hotel sales show a double peak, from April to June and August to October respectively. Resort hotels have the highest annual sales in July and August, showing a clear low and peak season.

Chart, bar chart

Description automatically generated

Figure 12 - Average Annual Room Revenue

From Figure 13 we can observe that City hotel prices fluctuate less throughout the year, while Resort hotel prices per night are more volatile with the month, with the highest prices in August. The best times to book city hotels and resorts are November and January of the year.

Chart, line chart

Description automatically generated

Figure 13 – Average room price per night

1. https://www.kaggle.com/jessemostipak/hotel-booking-demand [↑](#footnote-ref-2)
2. “5 Factors That Can Affect the Price of Your Hotel Rooms.” Insights, 29 Jan. 2020, https://insights.ehotelier.com/insights/2020/02/05/5-factors-that-can-affect-the-price-of-your-hotel-rooms/. [↑](#footnote-ref-3)
3. Falk, Martin & Vieru, Markku. (2018). Modelling the cancellation behaviour of hotel guests. International Journal of Contemporary Hospitality Management. 30. 10.1108/IJCHM-08-2017-0509. [↑](#footnote-ref-4)
4. McKay, Carol. “Hotels Holding Consumers Hostage with Increasingly Unfriendly Cancellation Policies - National Consumers League.” National Consumers League, Carol McKay Https://Nclnet.org/Wp-Content/Uploads/2020/08/NCL-Logo.png, 30 June 2021, https://nclnet.org/hotels\_cancellation\_policies/. [↑](#footnote-ref-5)