Task to solve:

**Multiple time series prediction of demand for each room category**

**Category in columns “ReservedRoomType” and “AssignedRoomType”: A, B, C,… etc. Need to predict demand for the ReservedRoomType**

**Free choice: model, quality metric, prediction period can cover from 1 week to 1 month (let you choose).**

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. 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.

**Specifications table**

|  |  |
| --- | --- |
| **Subject area** | **Hospitality Management** |
| **More specific subject area** | Revenue Management |
| **Type of data** | Text files (csv) |
| **How data was acquired** | Extraction from hotels’ Property Management System (PMS) SQL databases |
| **Data format** | Mixed (raw and preprocessed) |
| **Experimental factors** | Some of the variables were engineered from other variables from different database tables. The data point time for each observation was defined as the day prior to each booking׳s arrival |
| **Experimental features** | Data was extracted via TSQL queries executed directly in the hotels’ PMS databases and R was employed to perform data analysis |
| **Data source location** | Both hotels are located in Portugal: H1 at the resort region of Algarve and H2 at the city of Lisbon |

**Value of the data**

•

Descriptive analytics can be employed to further understand patterns, trends, and anomalies in data;

•

Used to perform research in different problems like: bookings cancellation prediction, customer segmentation, customer satiation, seasonality, among others;

•

Researchers can use the datasets to benchmark bookings’ prediction cancellation models against results already known ;

•

Machine learning researchers can use the datasets for benchmarking the performance of different algorithms for solving the same type of problem (classification, segmentation, or other);

•

Educators can use the datasets for machine learning classification or segmentation problems;

•

Educators can use the datasets to obtain either statistics or data mining training.

## 1. Data

In tourism and travel related industries, most of the research on Revenue Management demand forecasting and prediction problems employ data from the aviation industry, in the format known as the Passenger Name Record (PNR). This is a format developed by the aviation industry However, the remaining tourism and travel industries like hospitality, cruising, theme parks, etc., have different requirements and particularities that cannot be fully explored without industry׳s specific data. Hence, two hotel datasets with demand data are shared to help in overcoming this limitation.

The datasets now made available were collected aiming at the development of prediction models to classify a hotel booking׳s likelihood to be canceled. Nevertheless, due to the characteristics of the variables included in these datasets, their use goes beyond this cancellation prediction problem.

One of the most important properties in data for prediction models is not to promote leakage of future information. In order to prevent this from happening, the timestamp of the target variable must occur after the input variables’ timestamp. Thus, instead of directly extracting variables from the bookings database table, when available, the variables’ values were extracted from the bookings change log, with a timestamp relative to the day prior to arrival date (for all the bookings created before their arrival date).

Not all variables in these datasets come from the bookings or change log database tables. Some come from other tables, and some are engineered from different variables from different tables. A diagram presenting the PMS database tables from where variables were extracted is presented in [Fig. 1](https://www.sciencedirect.com/science/article/pii/S2352340918315191#f0005). A detailed description of each variable is offered in the following section.

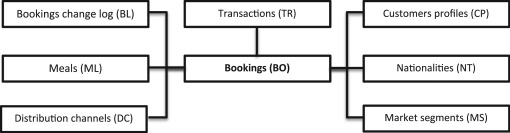


Fig. 1. Diagram of PMS database tables where variables where extracted from.

Table 1. Variables description.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Description** | **Source/Engineering** |
| ***ADR*** | Numeric | Average Daily Rate as defined by [[5]](https://www.sciencedirect.com/science/article/pii/S2352340918315191#bib5) | BO, BL and TR / Calculated by dividing the sum of all lodging transactions by the total number of staying nights |
| ***Adults*** | Integer | Number of adults | BO and BL |
| ***Agent*** | Categorical | ID of the travel agency that made the booking[a](https://www.sciencedirect.com/science/article/pii/S2352340918315191#tbl1fna) | BO and BL |
| ***ArrivalDateDayOfMonth*** | Integer | Day of the month of the arrival date | BO and BL |
| ***ArrivalDateMonth*** | Categorical | Month of arrival date with 12 categories: “January” to “December” | BO and BL |
| ***ArrivalDateWeekNumber*** | Integer | Week number of the arrival date | BO and BL |
| ***ArrivalDateYear*** | Integer | Year of arrival date | BO and BL |
| ***AssignedRoomType*** | Categorical | Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons | BO and BL |
| ***Babies*** | Integer | Number of babies | BO and BL |
| ***BookingChanges*** | Integer | 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 | BO and BL/Calculated by adding the number of unique iterations that change some of the booking attributes, namely: persons, arrival date, nights, reserved room type or meal |
| ***Children*** | Integer | Number of children | BO and BL/Sum of both payable and non-payable children |
| ***Company*** | Categorical | ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons | BO and BL. |
| ***Country*** | Categorical | Country of origin. Categories are represented in the ISO 3155–3:2013 format [[6]](https://www.sciencedirect.com/science/article/pii/S2352340918315191#bib6) | BO, BL and NT |
| ***CustomerType*** | Categorical | Type of booking, assuming one of four categories: | BO and BL |
| 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 |
| ***DaysInWaitingList*** | Integer | Number of days the booking was in the waiting list before it was confirmed to the customer | BO/Calculated by subtracting the date the booking was confirmed to the customer from the date the booking entered on the PMS |
| ***DepositType*** | Categorical | Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: | BO and TR/Value calculated based on the payments identified for the booking in the transaction (TR) table before the booking׳s arrival or cancellation date. |
| No Deposit – no deposit was made; |
| In case no payments were found the value is “No Deposit”. |
| If the payment was equal or exceeded the total cost of stay, the value is set as “Non Refund”. |
| Non Refund – a deposit was made in the value of the total stay cost; |
| Otherwise the value is set as “Refundable” |
| Refundable – a deposit was made with a value under the total cost of stay. |
| ***DistributionChannel*** | Categorical | Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators” | BO, BL and DC |
| ***IsCanceled*** | Categorical | Value indicating if the booking was canceled (1) or not (0) | BO |
| ***IsRepeatedGuest*** | Categorical | Value indicating if the booking name was from a repeated guest (1) or not (0) | BO, BL and C/ Variable created by verifying if a profile was associated with the booking customer. If so, and if the customer profile creation date was prior to the creation date for the booking on the PMS database it was assumed the booking was from a repeated guest |
| ***LeadTime*** | Integer | Number of days that elapsed between the entering date of the booking into the PMS and the arrival date | BO and BL/ Subtraction of the entering date from the arrival date |
| ***MarketSegment*** | Categorical | Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators” | BO, BL and MS |
| ***Meal*** | Categorical | Type of meal booked. Categories are presented in standard hospitality meal packages: | BO, BL and ML |
| 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) |
| ***PreviousBookingsNotCanceled*** | Integer | Number of previous bookings not cancelled by the customer prior to the current booking | BO and BL / In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and not canceled. |
| ***PreviousCancellations*** | Integer | Number of previous bookings that were cancelled by the customer prior to the current booking | BO and BL/ In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and canceled. |
| ***RequiredCardParkingSpaces*** | Integer | Number of car parking spaces required by the customer | BO and BL |
| ***ReservationStatus*** | Categorical | Reservation last status, assuming one of three categories: | BO |
| 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 |
| ***ReservationStatusDate*** | Date | Date at which the last status was set. This variable can be used in conjunction with the *ReservationStatus* to understand when was the booking canceled or when did the customer checked-out of the hotel | BO |
| ***ReservedRoomType*** | Categorical | Code of room type reserved. Code is presented instead of designation for anonymity reasons | BO and BL |
| ***StaysInWeekendNights*** | Integer | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel | BO and BL/ Calculated by counting the number of weekend nights from the total number of nights |
| ***StaysInWeekNights*** | Integer | Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel | BO and BL/Calculated by counting the number of week nights from the total number of nights |
| ***TotalOfSpecialRequests*** | Integer | Number of special requests made by the customer (e.g. twin bed or high floor) | BO and BL/Sum of all special requests |

a

ID is presented instead of designation for anonymity reasons.

The PMS assured no missing data exists in its database tables. However, in some categorical variables like Agent or Company, “NULL” is presented as one of the categories. This should not be considered a missing value, but rather as “not applicable”. For example, if a booking “Agent” is defined as “NULL” it means that the booking did not came from a travel agent.