# Appendixes

## Appendix A – Features description

| Feature | Type | Data source | Models | Description |
|---|---|---|---|---|
| Adults | N, I | P | 1-4 | Number of adults |
| Agent | C, I | P | 1-4 | ID of agency (if booked through an agency). The levels differ by hotel as each hotel has a list of agencies it works with. |
| AssociatedToEvent | C, E | P | 4 | Binary value indicating whether the booking was associated with an event held at the hotel (e.g., meeting or wedding) (0: no; 1: yes) |
| AvgQuantityOfPrecipitationInMM | N, E | W, P | 1-4 | Average quantity of precipitation forecasted. This value was calculated by summing the milliliters of precipitation forecast for the days of the stay and dividing that by the number of days of the stay for which there was a weather forecast. The values of the forecast were accounted according to the booking cancellation outcome. For bookings that were not canceled, the arrival date was considered. For canceled bookings, the cancellation date was considered |
| Babies | N, I | P | 1-4 | Number of babies |
| BookingChanges | N, E | P | 1-4 | Heuristic created by summing the number of booking changes (amendments) prior to arrival that could indicate cancellation intentions (arrival or departure dates, number of persons, type of meals, ADR, or reserved room type). Each variable change is counted as one change. For example, if the arrival date and number of persons were changed in a single operation, that would be counted as two changes |
| BookedSPA | C, E | P | 4 | Binary value indicating whether an SPA service was booked prior to the guest's arrival (0: no; 1: yes) |
| Children | N, I | P | 1-4 | Number of children |
| Company | C, I | P | 1-4 | ID of company/corporation (if an account was associated with it). The levels differ by hotel as each hotel has a list of companies it works with. |
| CompSetSocialReputationDifference | N, E | R | 1-4 | Number of hotels in the competition set that had a better rating booking outcome date (arrival or cancellation date according to the outcome). This feature value is obtained by summing the number of hotels from the competitive set that, at the booking outcome date, had better social reputation rating (normalized |

| Feature | Type | Data source | Models | Description |
|---|---|---|---|---|
| | | | | aggregated rating) than the tested hotel. |
| Country | C, I | P | 1-4 | Country ISO 3166 alpha-3 identification of the main booking holder |
| CustomerType | C, E | P | 1-4 | Type of customer (group, contract, transient, or transient-party); the last category is a heuristic built when the booking is transient but is fully or partially paid in conjunction with other bookings (e.g., small groups such as families who require more than one room) |
| DayOfYear | N, E | P | 1-4 | Number representing the sequential day of the year. For example, January 1st is 1, and February 1st is 32. |
| DaysInWaitingList | N, I | P | 1-4 | Number of days the booking was on a waiting list prior to confirmed availability and confirmation as a booking |
| DepositType | C, E | P | 1-4 | Since hotels had different cancellation and deposit policies, a heuristic was developed to define the deposit type (nonrefundable, refundable, no deposit): payment made in full before the arrival date was considered a "nonrefundable" deposit, and partial payment before arrival was considered a "refundable" deposit; otherwise it was considered as "no deposit". |
| DistributionChannel | C, I | P | 1-4 | Distribution channel used to make the booking (e.g., OTA, Direct, Travel Operator). The levels differ by hotel, as each hotel works with different distribution channels. |
| HotelsWithRoomsAvailable | N, E | O | 1-4 | Number of competitors that have inventory available for the period of the booking stay with the same type of meal package and that could accommodate the same number of adults. Inventory availability is obtained by checking the availability for all nights of the stay (lookup date) of all hotels in the competitive set, on the arrival or cancellation date according to the cancellation outcome (observation date) |
| IsRepeatedGuest | C, E | P | 1-4 | Binary value indicating whether the booking holder, at the time of booking creation, was a repeat guest at the hotel (0: no; 1: yes); created by comparing the time of booking with the guest profile creation record |
| LeadTime | N, E | P | 1-4 | Number of days prior to arrival that the hotel received the booking (usually, the date when the booking was entered in the PMS) |

| Feature | Type | Data source | Models | Description |
|---|---|---|---|---|
| MarketSegment | C, I | P | 1-4 | Market segment in which the booking was classified. The levels differ by hotel, as each hotel works with different market segments |
| Meal | C, I | P | 1-4 | ID of meal the guest requested. The levels differ by hotel as each hotel works with different types of meals. |
| nHolidays | N, E | H, P | 1-4 | Number of local holidays that are due to occur during the booking stay (including the check-out date) |
| PreviousCancellationRatio | N, E | P | 1-4 | Ratio calculated by dividing the guest's number of previous cancellations by the guest's previous number of bookings at the hotel (as of the booking creation date) |
| RatioADRbyCompsetMedianDifference | N, E | O, P | 1-4 | Ratio calculated by dividing the booking ADR by the average of the median price at the competitor hotels for the cheapest room each competitor had available that included the same type of meal package and could accommodate the number of adults in the booking. Competitors' prices are obtained on the arrival or cancellation date according to the cancellation outcome. In other words, the competitor's median prices for each stay night (lookup date) at the booking outcome date (observation date) are summed and divided by the number of nights to obtain a competitor's set average median price. A ratio is then calculated by dividing the booking ADR by the competitor's set median average price. The objective of this feature is to understand whether the competitors are offering a better or worse price and the amplitude of the difference. |
| RatioMajorEventsNights | N, E | S, P | 1-4 | Ratio calculated by dividing the total number of major special events that are expected to occur during the stay by the total number of nights of the booking |
| RatioMinorEventsNights | N, E | S, P | 1-4 | Ratio calculated by dividing the total number of minor special events that are expected to occur during the stay by the total number of nights of the booking |
| RequiredCarParkingSpaces | N, I | P | 4 | Number of car parking spaces required by the guest |
| ReservedRoomType | C, I | P | 1-4 | Room type requested by the guest |
| SRDoubleBed | C, E | P | 4 | Binary value indicating whether the guest, prior to arrival, asked specifically for a double bed (0: no; 1: yes) |
| SRHighFloor | C, E | P | 4 | Binary value indicating whether the guest, prior to arrival, asked |

| Feature | Type | Data source | Models | Description |
|---|---|---|---|---|
| | | | | specifically for a room on a high floor (0: no; 1: yes) |
| SRQuietRoom | C, E | P | 4 | Binary value indicating whether the guest, prior to arrival asked specifically for a quiet room (0: no; 1: yes) |
| SRTogether | C, E | P | 4 | Binary value indicating whether the guest, prior to arrival, asked specifically to be placed in a room close to another booking (0: no; 1: yes) |
| SRTwinBed | C, E | P | 4 | Binary value indicating whether the guest, prior to arrival, asked specifically for a twin bed (0: no; 1: yes) |
| StaysInWeekendNights | N, E | P | 1-4 | How many nights of the total stay were on weekends (Saturday and Sunday) |
| StaysInWeekNights | N, E | P | 1-4 | How many nights of the total stay were on weekdays (Monday through Friday) |
| ThirdQuartileDeviationADR | N, E | P | 1-4 | Ratio calculated by dividing the booking ADR by the third quartile value of all bookings for the same distribution channel and the same reserved room type in the same expected week/year of arrival. The objective of this feature is to understand whether the price of the booking is much higher than that of other similar bookings. |
| TotalOfSpecialRequests | N, E | P | 1-4 | Number of special requests made (e.g., fruit basket, sea view, etc.) |

Legend:

- Type:
  - C – Categorical
  - E – Engineered
  - I – Input
  - N – Numeric
- Data source:
  - H – Holiday calendar
  - O – Online prices/inventory
  - P – PMS
  - R – Social media reputation
  - S – Special events calendar
  - W – Weather

## Appendix B – Machine learning metrics

**Accuracy (Acc.)**: Measure of outcome correctness; it measures the proportion of true results among the total number of predictions. The formula is as follows: $Acc. = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN}$.

**Area Under the Curve (AUC)**: Measure of success calculated from the area under the plot of true positive rate (TPR) against false positive rate (FPR).

**False Negative (FN)**: The outcome prediction was negative, but the actual value was positive (e.g., the booking was predicted as likely not to cancel, but it was canceled).

**False Positive (FP)**: The outcome prediction was positive, but the actual value was negative (e.g., the booking was predicted as likely to cancel, but it was not canceled).

**False Positive Rate (FPR or Fall-out)**: Measures the probability of a positive prediction result and the actual value being negative (e.g., the probability of a booking being predicted as likely to cancel and effectively did not cancel). The formula is as follows: $FPR = \frac{\sum FP}{\sum FP + \sum TN}$.

**Precision (Pre.)**: Measures the proportion of correct positive predictions. The formula is as follows: $Pre. = \frac{\sum TP}{\sum TP + \sum FP}$.

**True Negative (TN)**: The outcome prediction was negative, and so was the actual value (e.g., the booking was predicted as likely not to cancel and was not canceled).

**True Positive (TP)**: The outcome prediction was positive, and so was the actual value (e.g., the booking was predicted as likely to cancel and was canceled).

**True Positive Rate (TPR or Sensitivity)**: Measures the probability of a positive prediction result and the actual value being positive (e.g., the probability of a booking being identified as likely to cancel and being canceled). The formula is as follows: $TPR = \frac{\sum TP}{\sum TP + \sum FN}$.