MSiA 401: Mini Project - Group 6

Alejandra Lelo de Larrea Ibarra Kiran Jyothi Sheena Lixuan Chen
Wencheng Zhang

Executive Summary

HERE GOES THE EXECUTIVE SUMMARY

Contents

1 Introduction

Smith Travel Research estimates there are 17.5 million guestrooms in 187,000 hotels worldwide. In 2020, global hotel revenue was $198.6 billion dollars, and the hotel and tourism industry accounts for approximately 10% of worldwide GDP (Hollander Accessed March 3rd 2023). The hotel industry experiences approximately 24% of cancellations on reservations, and this rate increases up to 38% for online bookings (Loeb Accessed March 3rd 2023). With this figures, and specially after the Covid-19 Pandemic, the study of hotel booking cancellations has become more relevant.

Accurately forecasting hotel booking cancellations and understanding the factors that influence such behavior is relevant for the following XXX reasons. First, from an economic perspective, each unoccupied room results in an economic loss for the hotel. If accurately and timely forecasted, the hotel could still allocate the room to a different customer and avoid loosing revenue. Additionally, this would help the hotel to understand their net demand. Second, from the operational side of the business, managing cancellations is costly both in time and resources. For example, during peak seasons the hotel might need to allocate additional resources to manage cancellations and to try to reallocate the rooms. An accurate forecast could help to reduce operational costs and improve the hotel's efficiency.

Third, from a marketing perspective, forecasting booking cancellations opens the possibility to implement pricing strategies, such as offering discounts to retain the customer.

Hotel booking cancellations and revenue management has been widely studied in the literature. (Antonio, Almeida, and Nunes 2019) use data from eight hotels combined with additional sources (such as weather or holidays) to develop booking cancellation prediction models to help hoteliers understand their net demand and improve their revenue management. (Chen, Schwartz, and Vargas 2011) use a multinomial logit regression to analyze the impact of cancellation fees and deadlines on hotel bookings and find that the former affects customer's behavior but the later doesn't. (Falk and Vieru 2018) uses data of 9 hotels to estimate a probit model with cluster adjusted standard errors at the hotel level to find the determinants of cancellation probability, among which booking lead time and country of residence are the most important. (Sanchez-Medina, Eleazar, et al. 2020) apply ML algorithms to forecast hotel booking cancellations using genetic algorithms to configure the structural parameters of the artificial neural network used.

In this regard, we contribute to the study of hotel booking cancellations by analyzing the *Hotel booking demand* data set available in Kagle[1] with the aim of forecasting hotel booking cancellations through the lens of Machine Learning (ML) models. This data set contains 19,390 booking registries of a resort hotel and a city hotel from a chain in Portugal and includes 32 predictors such as date of the booking, length of stay, number of children and adults, type of meal, number of special requests, among others. To forecast cancellations, we first do an exploratory data analysis to understand the data set and the relations between the predictors and the response. Then we clean the data, select the main features and apply some transformations, as well as create new features. Then we fit several ML classification models to the final data set. For this process, we randomly split the data in training (XXX observations) and test (XXXX observations). Specifically, we estimate a Naive Bayes classifier (benchmark), Logistic Regression, Neural Network, Tree, k-Nearest Neighboors, Local Logistic Regression, Weighted Kernel, Boosting, Random Forest and Support Vector Machines. For each of this models, we use 10-fold cross validation (CV) for hyperparameter tuning. Note that each model was run independently in this step. After finding the optimal hyperparameters, we apply again 10-fold cross validation to compare the optimal models. Finally, we evaluate the model performance in the test set. To compare models we use the miss-classification rate, the AUC measure and the F1-score. We find that the benchmark model (Naive Bayes) has a DESCRIBE RESULTS HERE....

The rest of the report is structured as follows. Section 2 presents the original data, exploratory data analysis, data cleaning and feature engineering strategy, and details the final data set to be used. Then, Section 3 succinctly presents each of the ML models selected to fit the data and the evaluation process. After this, Section 4 compares the forecasting power of each model. Lastly, Section 5 presents our final remarks.

2 Data

2.1 Original Data

Analyzing the *Hotel booking demand* data set available in KaggleThe raw dataset comprises 32 columns and 119,390 rows, with 12 columns being categorical and the rest numerical. Each row represents a unique booking record, containing booking dates, length of stay, number of guests, booking platforms, meal plan types, and other information. Table 2.1 details each variable in the dataset.

[1]Data set retrieve on January 18th 2023 at https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand.

Table 1: Data Description

Variable	Date Type	Description
ADR	*Numeric*	Average Daily Rate
Adults	*Integer*	Number of adults
Agents	*Categorical*	ID of the travel agency that made the booking
ArrivalDateDayOfMonth	*Integer*	Day of the month of the arrival date
ArrivalDateMonth	*Categorical*	Month of arrival date with 12 categories: "January" to "December"
ArrivalDateWeekNumber	*Integer*	Week number of the arrival date
ArrivalDateYear	*Integer*	Year of arrival date
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
Babies	*Integer*	Number of babies
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
Children	*Integer*	Number of 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
Country	*Categorical*	Country of origin. ISO 3155-3:2013 format
CustomerType	*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
DaysInWaitingList	*Integer*	Number of days the booking was confirmed
DepositType	*Categorical*	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit; Non Refund; Refundable
DistributionChannel	*Categorical*	Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"
IsCanceled	*Categorical*	Value indicating if the booking was BO canceled (1) or not (0)
IsRepeatedGuest	*Categorical*	Value indicating if the booking name was from a repeated guest (1) or not (0)
LeadTime	*Integer*	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
MarketSegment	*Categorical*	Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
Meal	*Categorical*	Type of meal booked. Categories are presented in standard hospitality meal packages: 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 canceled by the customer prior to the current booking
PreviousCancellations	*Integer*	Number of previous bookings that were canceled by the customer prior to the current booking
RequiredCardParkingSpaces	*Integer*	Number of car parking spaces required by the customer
ReservationStatus	*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
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
ReservedRoomType	*Categorical*	Code of room type reserved. Code is presented instead of designation for anonymity reasons
StaysInWeekendNights	*Integer*	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
StaysInWeekNights	*Integer*	Number of week nights (Monday to Fri-day) the guest stayed or booked to stay at the hotel
TotalOfSpecialRequests	*Integer*	Number of special requests made by the customer (e.g. twin bed or high floor)

To explore the data, we calculated descriptive statistics and created visualizations to illustrate the structure and distribution of our dataset. Four of the columns have missing values. The "children" column has four missing values, while the "country" column has 488 missing values. The "agent" column has 16,340 missing values, accounting for around 14% of the total rows. The "company" column has 112,593 missing values, indicating that the majority of rows lack company information.

It is important to note that some categorical variables in the data have an excessively large number of classes or categories. One-hot encoding these variables would result in an unwieldy and expensive data set with a vast number of dimensions. Table 2 shows the number of categories in each variable.

Table 2: Number of Categories per Variable

Categorical Variable	**# of Categories**
arrival_date_month	12
assigned_room_type	12
country	177
customer_type	4
deposit_type	3
distribution_channel	5
hotel	2
market_segment	8
meal	5
reservation_status	3
reservation_status_date	926
reserved_room_type	10

2.2 EDA

Upon cleaning and one-hot encoding the raw data, we have the following findings:

1. We didn't find strong correlation between the variables (see Figure 1).

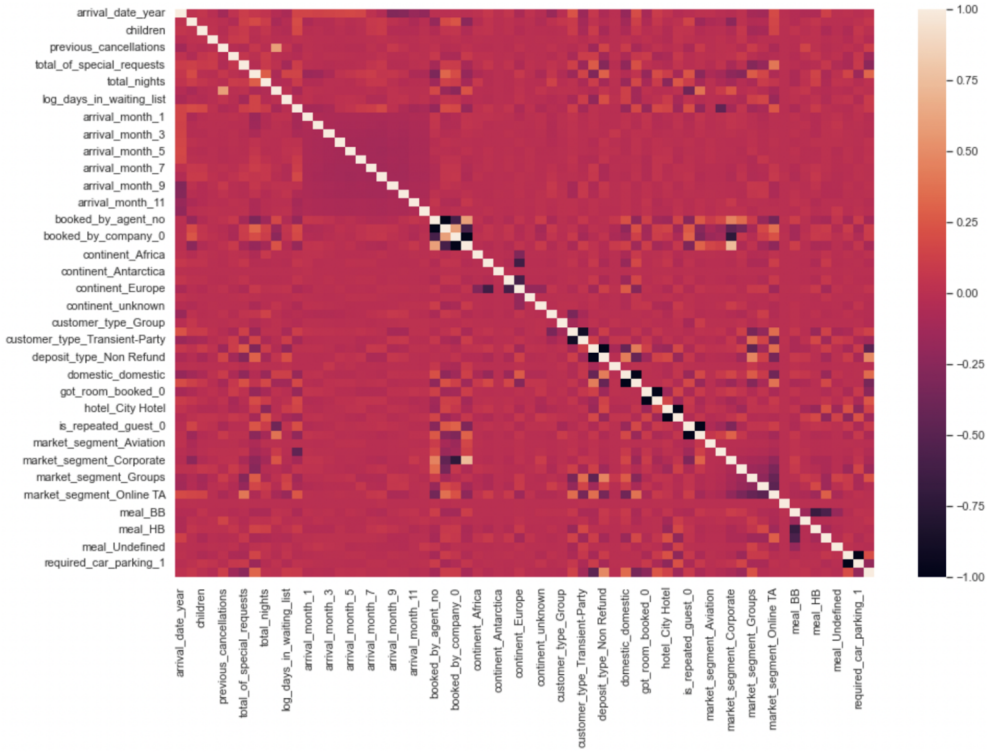

Figure 1: Correlation Plot

2. The percentage of repeated guests that cancel reservations (17%) is considerably lower than the percentage of first-time customers that end up canceling the hotel booking (61%).

3. For guests that opt for a non refundable booking, there is a surprisingly high proportion of cancellations in comparison to non cancellations. This phenomenon appears to challenge the notion that the absence of a refund would generally discourage cancellations. (See Table 3)

Table 3: Number of Categories per Variable

is_canceled	No Deposit	Non Refund	Refundable
0	74947	93	126
1	29694	14494	36

4. A large majority of the guests are of Portuguese descent which matches general intuition considering the notion that the data reflect hotels based in Portugal. Furthermore, other common countries of origin are mostly from neighboring countries including the UK, France, and Spain (see Figure 2).

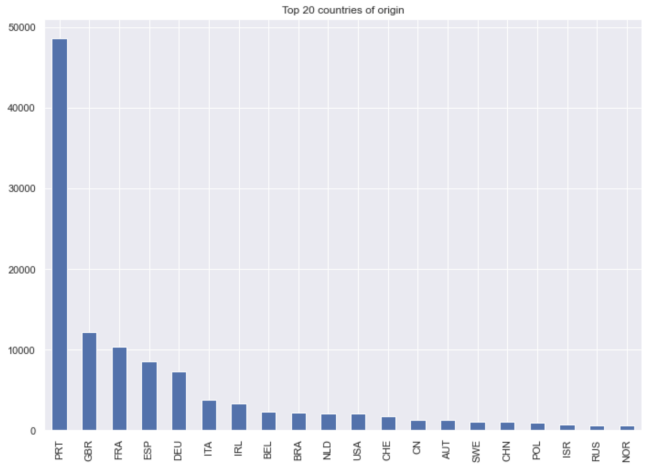

Figure 2: Top 20 Countries of Origin

5. Individual and family stays constitute a large majority of the data; on the other hand, contract and group reservations appear to be far less prominent. (see Figure 3)

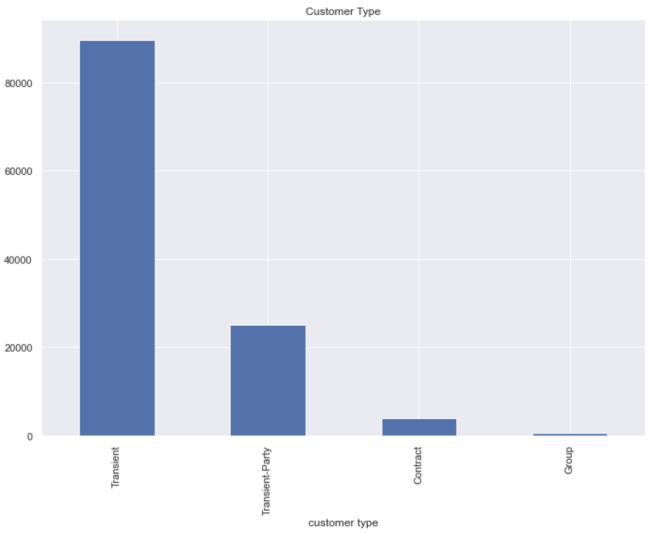

Figure 3: Customer Type Distribution

6. Bookings with a considerably large number of adults all appear to have been canceled. The graph also hints at potential outliers in the count of adults included in the reservation (see Figure 4).

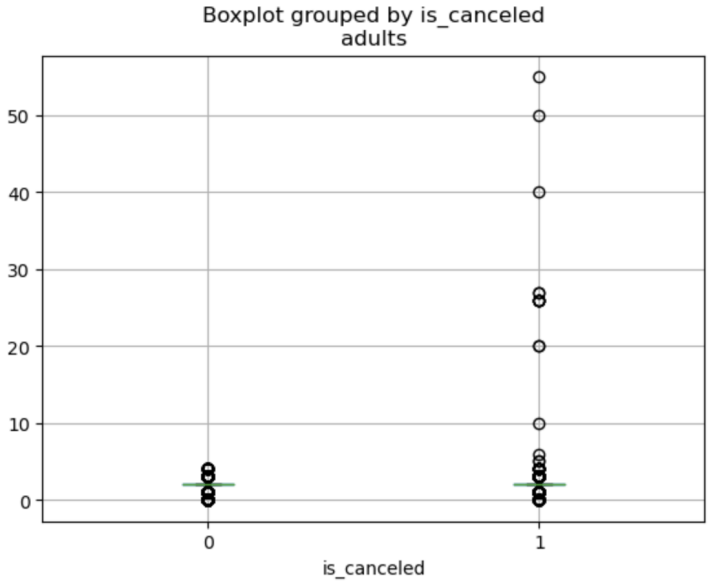

Figure 4: Adult distribution by cancellation.

7. The number of days a reservation is held in the waiting list is typically not too long and appears reasonable from a transactional standpoint. However, there are instances where the number of days a reservation was held in the waiting list prior to confirmation was considerably long (see Figure 5)

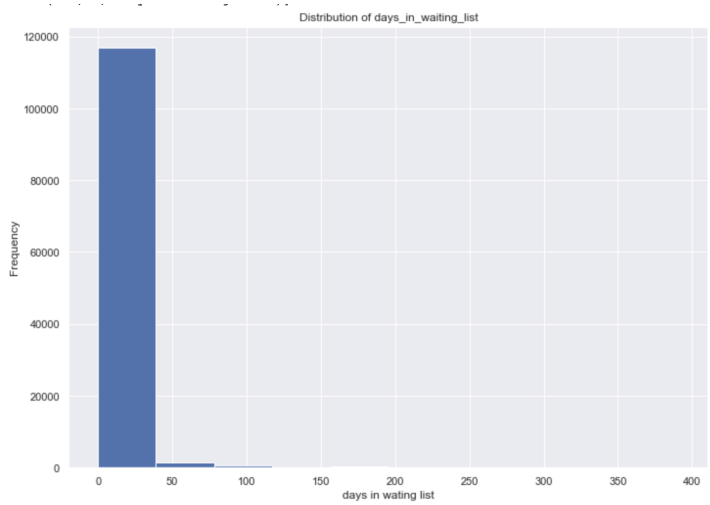

Figure 5: Days in waiting list.

8. Aggregating the bookings by arrival date, a clear pattern of seasonality can be observed from the data: the demand for hotel bookings surges during the summers and winters (see Figure 6 and Figure 7).

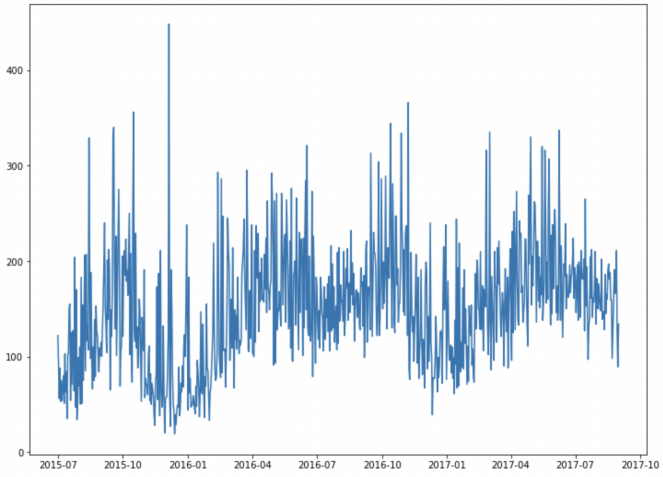

Figure 6: Number of Bookings for each Arrival Date.

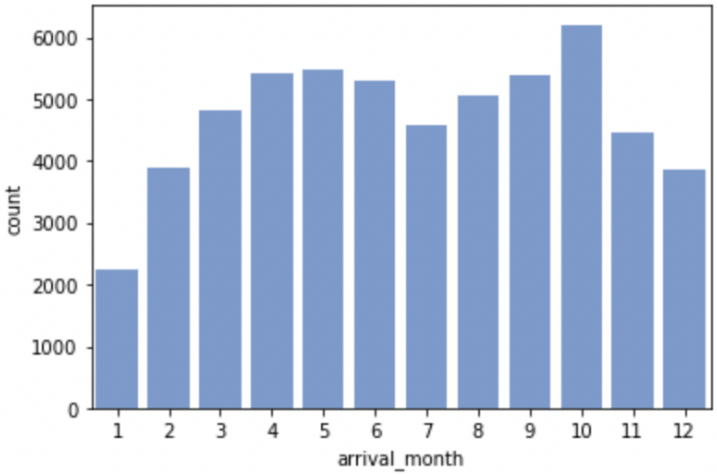

Figure 7: Number of Bookings for each Arrival Month

9. The duration of stays variables are highly right-skewed. From Table 4 we can see that there are some bookings with very large outliers, while the majority of the stays last between 1 and 5 days. If we bin the columns, we observe roughly 46 rows with stays greater than 20 days.

Table 4: Length of stay in days.

Class	**Count**
0-5	114537
6-10	4451
11-15	245
16.20	111
>20	46

10. There does not appear to be a meaningful difference in the mean number of stays and composition of guests between canceled and not canceled reservations (see Table 5).

Table 5: Number of stays in week days and guest composition.

is_canceled	Stays in week nights	Adults	Children	Babies
0	2.4640	1.8297	0.1023	0.0103
1	2.5619	1.9017	0.1065	0.0038

2.3 Data Cleaning, Transformations & Feature Engineering

2.4 Final Data Set Description

3 Modelling

3.1 Train, Test & Cross Validation

To be able to buld and test the predictive power of the different models, we randomly divided the data into train and test sets making sure to keep the proportion of cancellations in both of them. The train set has 95,511 observations and the test set has a total of 23,877 observations. For each model, we use 10-fold cross validation to find the optimal hyperparameters (if needed), we then retrain the optimal model using the entire training set and asses its predictive power using the test set. To assess the effectiveness of each model, we used accuracy, F-1 score, AUC score and ROC.

3.2 Naive Bayes (benchmark)

3.3 Logistic Regression

3.4 Neural Networks

3.5 Tree

3.6 Nearest Neighboors

3.7 Local Logistic Regression

3.8 Weighted Kernel

3.9 Boosting

3.10 Random Forest

3.11 Support Vector Machines

4 Model Comparison

5 Final Remarks

References

Antonio, Nuno, Ana de Almeida, and Luis Nunes. 2019. "Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights into Booking Cancellation Behavior." *Cornell Hospitality Quarterly* 60 (4): 298–319.

Chen, Chih-Chien, Zvi Schwartz, and Patrick Vargas. 2011. "The Search for the Best Deal: How Hotel Cancellation Policies Affect the Search and Booking Decisions of Deal-Seeking Customers." *International Journal of Hospitality Management* 30 (1): 129–35.

Falk, Martin, and Markku Vieru. 2018. "Modelling the Cancellation Behaviour of Hotel Guests." *International Journal of Contemporary Hospitality Management* 30 (10): 3100–3116.

Hollander, Jordan. Accessed March 3rd 2023. "50+ Hospitality Statistics You Should Know (2023)." Hotel Tech Report; https://hoteltechreport.com/news/hospitality-statistics.

Loeb, Tony. Accessed March 3rd 2023. "Where Do Cancellations Come From?" Hotel Tech Report; https://hoteltechreport.com/news/hospitality-statistics.

Sanchez-Medina, Agustin J, C Eleazar, et al. 2020. "Using Machine Learning and Big Data for Efficient Forecasting of Hotel Booking Cancellations." *International Journal of Hospitality Management* 89: 102546.