Data Science in practice MGT-415

# PREDICTING HOTEL BOOKING CANCELLATION WITH MACHINE LEARNING

May 11, 2020

Jingwei CHEN
Ju-Hsuan HSIEH
Junze LI
Yuening ZHAN
Yunbei WANG

# Contents

# 1 Introduction

## 1.1 Background

In the hospitality industry, which means distributing rooms in a hotel, the revenue management is defined as 'making the right room available for the right guest and the right price at the right time via the right distribution channel' (Mehrotra & Ruttley, 2006, p. 2). As the biggest part of the hotel asset is fixed, hotel normally allow customers to make the room reservation in advance, in order to get the early paid cash flow and assign the rooms correctly to the 'right consumers'. (Talluri & Van Ryzin, 2004). This early payment, which separates the time of booking behavior and consuming behavior, has gradually become more and more standardized in hotels and other travel-related business. (Moe and Fader, 2002; Shugan and Xie, 2005).

Nevertheless, hotel rooms booking not only provides consumers with convenience when making travel plan, but also exposes them under the risk of lower future price. And the earlier the reservations before the checking in date, the bigger possibility there is. (Schwartz, 2000,2006,2008).

As a result, when making the travel plan, transient customers usually looking through many different hotels and even the same hotels through different booking agents, in order to get the best dealer. And there are more and more online tips shared by the savvy travelers to teach others how to save money on hotel rooms. For example, a pattern of cancellation called 'juggling', in which customers reserve rooms for multiple hotels, then choose one room/hotel and cancel the others at the last minute (Mandelbaum, 2016)The wide implement of online booking system unexpectedly increases the amount of the last-minute cancellation, because of the convenience of booking and cancelling hotel rooms. (Takizawa, 2017).

As the using rights of hotel rooms is regarded as 'perishable product', this behavior of last-minutes cancellation and no-show, (without cancellation) cause particularly high loss to the hotels, since there is no enough time to sell the 'unwanted rooms' to other people in need. (Xie and Gerstner, 2007; Koide and Ishii, 2005).

In order to solve the problem associated to the last-minute cancellation, there are two most commonly used strategies, the first one is overbooking and the second one is cancellation refund policies. (C.-C. Chen et al., 2011; C.-C. Chen & Xie, 2013)

Overbooking is that the hotels would let the same rooms booked by multiple customers, and get more orders compared with the actual amount of existing rooms. Although it is shown to reduce the risk of customers making orders that are not honored. (DeKay et al., 2004). But unfortunately, there are much bigger risk hidden behind. According to researches, overbooking strategy has potentially negative impact on customers satisfaction, especially when there appears more than one customer acclaim the ownership of the same room. The customers loyalty and future booking behavior would also be affected consequently, which causes big loss to the future revenues. (Lindenmeier and Tscheulin, 2008; Wangenheim and Bayón, 2007).

As for the cancellation refund policies, in terms of how much the tourism industry charges for cancellations, these penalties range from 10% or less to 100% of the rate (Phillips, 2005, p. 236). The general rule is, the closer the cancellation is to the day of check-in, the less percentage of the charged fee can be refunded to the account of customers. Most hotels (over 80% according to Engle, 2009) now charge a late cancellation penalty. One unexpected result is that hotel even gain profit from their cancellation refund strategies. In 2001, hotel cancellation fees represented 0.2% of the industry revenue (De Lollis, 2002). A newer report shows that this number grows to be 6.3%, from only the group cancellations (Mandelbaum, 2010). Such cancellation policies mitigate the revenues lost because of cancellations by capturing penalties (DeKay et al., 2004).

However, these policies would increase the travel budget of some unwary customers, so that are usually described as the 'hidden traps' by the annoyed customers. As the result, different

cancellation refund policy would be regarded as influential aspect to the deal-seeking consumers when making advanced booking. (Perkins, 2004).

## 1.2 Model Description

To overcome the negative impact caused by overbooking and the implementation of rigid cancellation policies to cope with booking cancellations, that can represent up to 20% of the total bookings received by hotels (Morales & Wang, 2010). This study is aimed at building a booking cancellation prediction model through data science, particularly, the classification in machine learning.

Similar hotel room cancellation predicting model has been developed by Chiang et al. (2007) and Antonio, Almeida and Luis Nunes (2017). But this study would apply different technique, programming language and analysis angle to data set of different hotels.

In this study, classification is used, because the data set contains a mixture of categorical and numerical variables, and the outcome would mostly be category. Firstly, this classification of detecting possible reservation cancellation is possible, as long as the suitable machine learning prediction algorithms are chosen. The other reason of choosing classification model, is that it is able to get not only categorical outcome as well as the numerical ones.

## 1.3 Data Description

In travel related industries, most of the research on Revenue Management focusing on forecasting issue, use the format known as the Passenger Name Record (PNR). Thanks to Antonio, Almeida and Nunes (2019). Besides the four hotels data used in their former study in 2017, they also collected the more updated data for another 2 hotels:40,060 observations for Hotel 1 and 79,330 observation for Hotel 2. This data set was obtained by them from the hotels' PMS databases' servers through the SQL Server Studio Manager.

The target of collecting this data set is to satisfy the demand of machine learning classification algorithm and can support to predicting the likelihood of hotel room cancellation. The variables' values were extracted from the bookings change log, with a timestamp relative to the day prior to arrival date. (Antonio, Almeida & Nunes, 2018)

There are 31 variables of 3 different types: *ReservationStatusDate* is the only date variable, 14 variables are categorical variables, the other 16 variables are numeric variables.

There is no missing data in this data set. For 'NULL' appearing in some categorical variables like *Agent* and *Company*, they can be understood as the customers make the reservations by themselves and through no travel agent. As for the other numbers in *Agent* and *Company* represent for the ID of corresponding travel agents and companies. *Iscancelled* and *IsRepeatedGuest* are dummy variables. There are four types of hotel rooms, which can be seen in both *AssignedRoomType* and *ReservedRoomType*. Normally, the hotel would assign the same type of room as the customer reserved. But sometimes the hotel has no asked room type, and then type codes under these two variables for one customer can be different. *Country* shows where the customers come from, and its categories are represented in the ISO 3155-3:2013 format.

*CustomerType* is one of the most important categorical variables, with 4 categories: 'Contract', 'Group', 'Transient' and 'Transient-party'. 'Transient' appears when the order is neither long-time contract nor part of group, and 'Transient-party' is 'Transients' orders associated with other transient booking. *DepositType* is another essential one which the customers would care most about, with 3 categories of 'No Deposit', 'Non-Refund' and 'Refundable'. Meanwhile, *ReservationStatus* is cared most by the hotel, also with 3 categories :'Cancelled' , 'Check-out' and 'No-show'.

*DistributionChannel* and *MarketSegment* shares similar categories: 'TA' means travel agents and 'TO' means tour operators. Meal categories are represented in standard hospitality meal packages. 'Und' is no meal, 'BB' is bed and breakfast, 'HB' is breakfast with one other meal and 'FB' is full board with all three meals.

The other 16 numerical variables are much easier to interpret. Most are straightforward date numbers, customer numbers or count numbers. The date numerical variables include *ArrivalDateWeekNumber*, *ArrivalDateOfMonth* and *ArrivalDateYear* ; *StaysInWeekendNights* and *StaysInWeekNights*, as well as *Leadtime* showing advanced booking period and *DaysInWaitinglist* that is usually less then 1 day. The date numerical variables contain *Adults*, *Babies* and *Children*. The count numerical variables are mostly related with booking and request behaviors, which are *BookingChanges*, *PreviousBookingsNotCancelled* and *PreviousBookingsCancellation*, as well as *RequiredCarParkingSpaces* and *TotalOfSpecialRequest*. Especially, the *ADR* is calculated by dividing the sum of all lodging transactions by the total number of staying nights.

Diagram explaining the relationship of variables from this data set is shown as Figure 1.

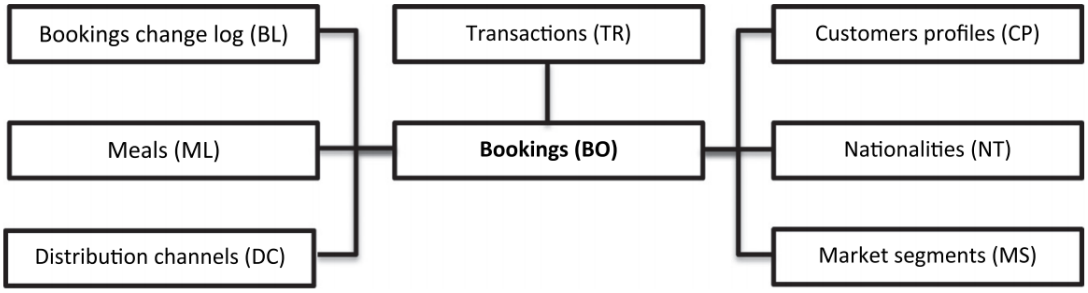



Figure 1: Diagram of PMS database tables where variables where extracted from.

# 2 Exploratory Data Analysis

## 2.1 Data Cleaning

The dataset contains 119390 booking records and each record contains 32 attributes. Firstly, we check if there are any missing values in the data set. We notice that there are some missing values in the attributes 'children', 'country', 'agent', and 'company'. For 'children' and 'country', there are only 4 and 488 missing values respectively, which are very small proportions in the dataset. So we drop the booking records which contain missing values in 'children' and 'country'. However, there are 16004 NULL values in the attribute 'agent' and 112275 in 'company'. As described above, the NULL means that the booking record is not made by agent/company. And when the booking is made by agent/company, there is ID of agent/company in the booking record. So we should not clean these two attributes directly right now, and we will handle these two attributes in the following part. After data cleaning, there are 118898 booking records remaining.

## 2.2 Data Analysis

First, this data set consists of Resort hotel and City hotel booking records, with 33% of Resort hotel data and 66% City hotel data. In the Resort hotel data, there are around 72% transactions not cancelled and 28% transactions cancelled. While in the City hotel data, there are around 58% transactions not cancelled and 41% transactions cancelled. In the overall data set, there are around 63% non-cancelled data and 37% cancelled data. Within all the cancelled data, there are 75% bookings are from City hotel.

Next, we look into each attributes to find the difference in value or distribution between cancelled and non-cancelled booking records. In the following parts, we highlight several interesting graphs that show the difference in these two types of bookings.

### 2.2.1 Analysing customer's country of origin

The country of origin where the customers come from is shown in Figure 2. Here we group them into continents based on the country code. Note that the y-axis is in logarithm scale because we find that the number of booking records is 90% from EU (Europe). This is because both the Resort hotels and the City hotels are located in Portugal, which is in Europe. In addition, we found that there is only non-cancelled booking data coming from Antarctica, where nobody lives there. There are only 2 transactions that came from Antarctica. There are several possibilities. One is that the data has an error, or that someone from other continent uses the IP from Antarctica.

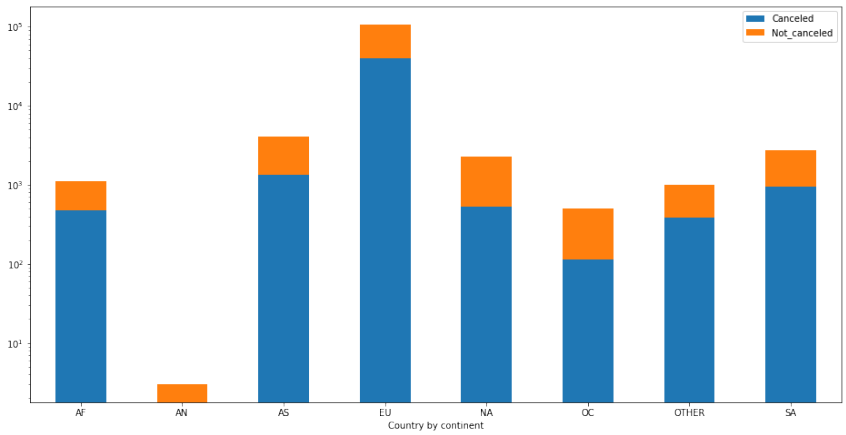



Figure 2: The country of origin of customers. AF:Africa, AN:Antarctica, AS:Asia, EU:Europe, NA:North America, OC:Oceania, SA: South America

### 2.2.2 Analysing the time between booking and arrival

The times between the date of booking and the date of arrival is defined as lead time. From the box plot of 'lead_time' (Figure 3), we can see that the distribution of the lead time of each transaction between cancelled and non-cancelled ones is different, with the cancelled ones being more concentrated around a higher value than the non-cancelled ones, meaning for the cancelled transactions, they booked it from an earlier date. That is to say, if the customer is booking the room at an early date, the possibility of canceling the booking is higher. In addition, we found a similar distribution when looking only at Resort hotel data or City hotel data. Therefore, for both Resort hotel or City hotel, the cancellation comes from an early data booking.

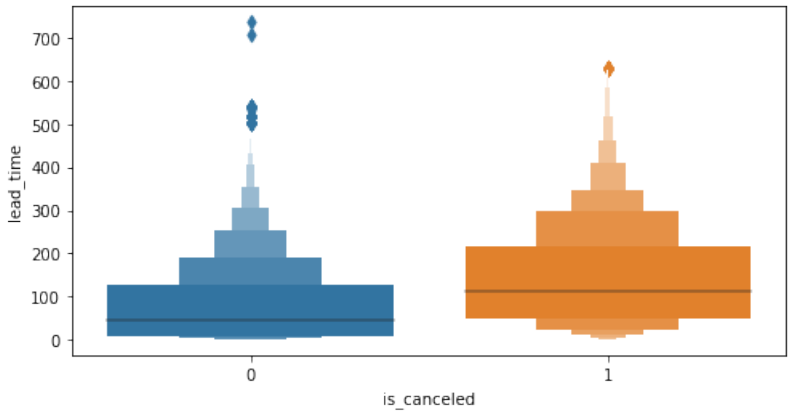



Figure 3: The lead time distribution of cancelled and non-cancelled data

### 2.2.3 Analysing the number of customers in a single booking record

The box plot below (Figure 4) indicated the number of adults in the booking record. We can see that for the transactions that are cancelled, a lot of them indicate that the order is for several adults, with the possibility of up to 50 people. In other words, a lot of these transactions are groups of people. We found that this phenomenon appears only in the bookings from Resort hotel. There is not much difference in the distribution in the City hotel booking record. This implies that how many adults are coming affect the cancellation more for the Resort hotel.

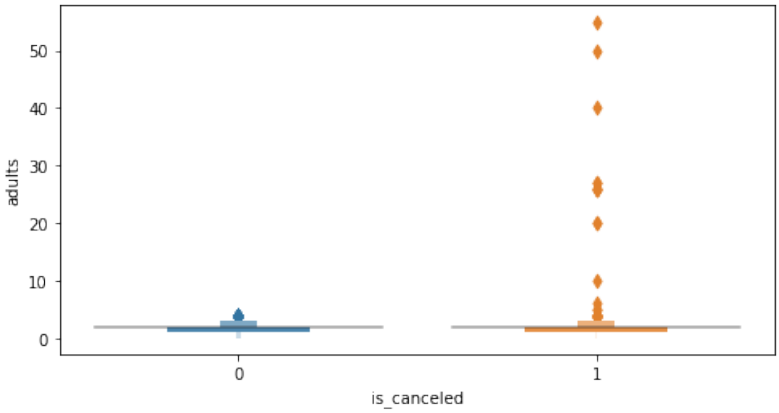



Figure 4: The # of adult distribution of cancelled and non-cancelled data

Moreover, we have also look into the distribution of the number of children of cancelled and non-cancelled data for both Resort hotel and City hotel booking records. We found that for both

Resort hotel and City hotels, there is not much difference between cancelled and non-cancelled transactions. This indicates that bringing children or not does not impact the cancellation.

### 2.2.4 Analysing customer's previous cancellation history

The box plot below (Figure5) indicates the number of previous cancellations of a certain customer. We can see that for the transactions that are not cancelled, there are only some data points that are non-zero (542 points). While for the transactions that are cancelled, there are more data points that are non-zero (5942 points), but concentrated at a low value (mainly 1). That is to say, for the current cancelled transaction, there are a lot of them that is previously cancelled once. In addition, when we looked into Resort hotel and City hotel data separately, we found that this phenomenon also holds. This implies that for both Resort hotel and City hotel, if this customer had cancelled once in the past, he is more likely to cancel the current booking as well.

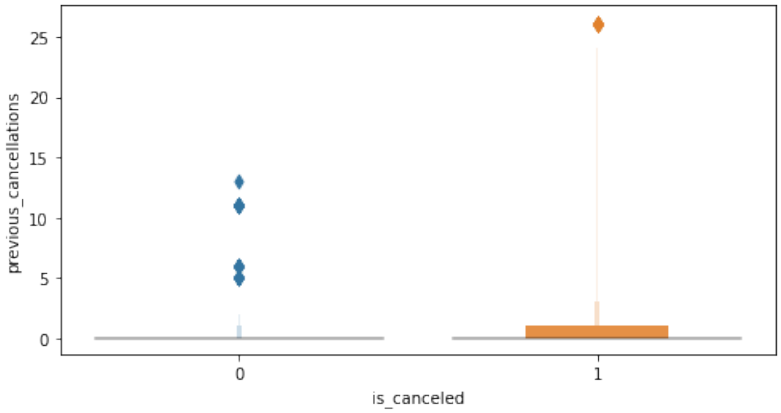



Figure 5: The # of previous cancellations distribution of cancelled and non-cancelled data

The box plot below (Figure 6) indicates the number of previous bookings that are not cancelled for this customer. For the transactions that are not cancelled, there are 3231 data points having non-zero value, and most of them have a higher value than the cancelled ones. While for the transactions that are cancelled, only 200 data points have non-zero values. That implies that for the customer that is currently not cancelled, they have more bookings that were previously not cancelled as well. Therefore, if a customer has a lot of bookings that are previously not cancelled, it is more possible that they are not going to cancel the current booking.

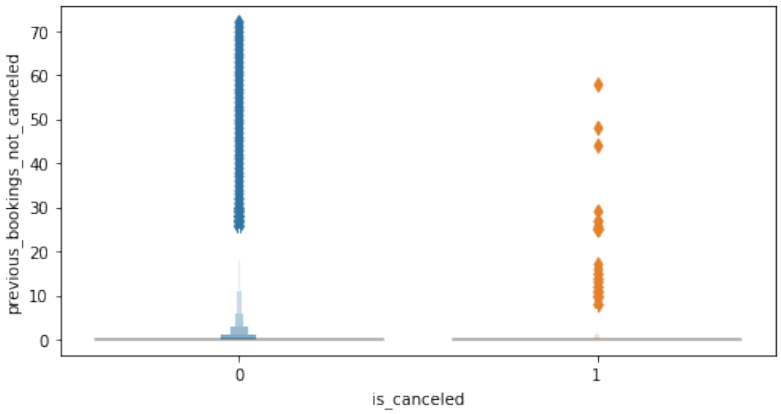



Figure 6: The # of previous bookings that are not cancelled distribution of cancelled and non-cancelled data

### 2.2.5 Analysing customer's booking changes

The box plot below (Figure 7) indicates the number of booking changes that the customer has done before the arrival date. We found that for the transactions that are not cancelled, there are 15243 data points having non-zero 'bookings_changes' values, while for the transactions that are cancelled, there are only 2833 having non-zero 'bookings_changes' values. That is to say, for the customers that did not finally cancel their bookings, they gad made more booking changes. This implies that the more booking changes they have done, they probably put more effort into the trip and thus they are not going to cancel their hotel bookings. We found this phenomenon both in Resort hotel and City hotel.

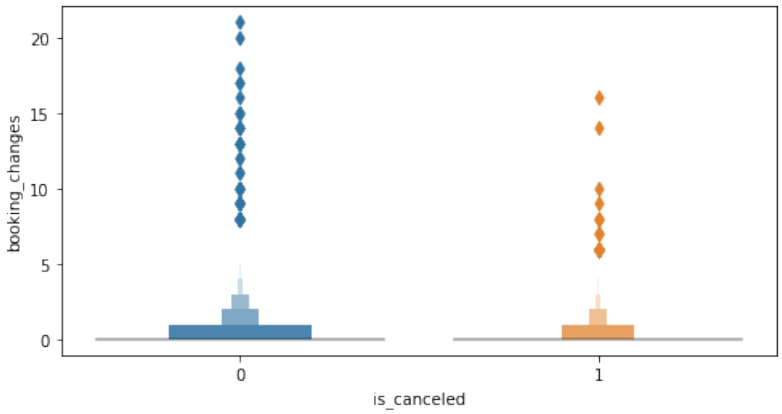



Figure 7: The # of booking changes of cancelled and non-cancelled data

### 2.2.6 Analysing customer's market segment

The bar chart below (Figure 8) show the market segment of the customer. We can see that the main market of both cancelled and non-cancelled transactions is an online travel agent. However, for the one that is cancelled, the second-highest market is from group customers then offline travel agent/tour operator and direct customers are far less than the two. While for the transactions that are not cancelled, the second highest is offline travel agent/tour operator than comes direct customers. We found this phenomenon in both Resort hotel and City hotel. This implies that group customers are more likely to cancel the bookings than offline travel agent/tour operator. In addition, for the customer that comes directly to book the hotel, they are no likely to cancel their bookings.

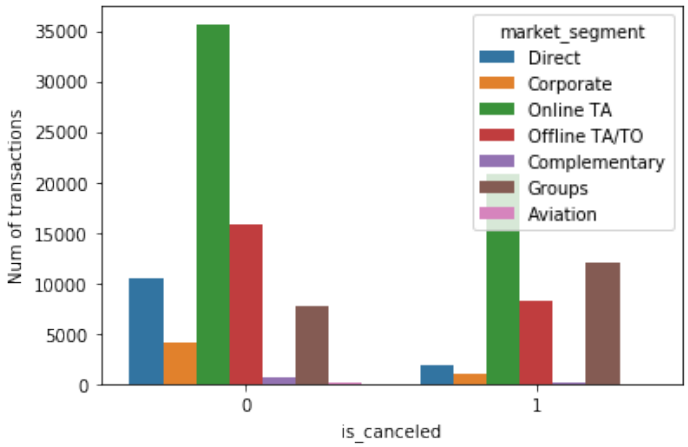



Figure 8: The market segment of cancelled and non-cancelled data

### 2.2.7 Analysing customers' special requests

The joint histogram shows the distribution of the bookings with different total special requests of cancelled and non-cancelled transactions (Figure 10). We can see that generally, the non-cancelled orders have more special requests than the cancelled ones. Moreover, the non-cancelled one decreases not as much as the cancelled one does. This indicates that the cancelled transactions generally have less special requests. Moreover, this might also imply that the more special requests the customer has, they probably put more effort into the trip, look forward to the trip and wish it was perfect. Thus, they are not going to cancel their hotel bookings. We found this phenomenon in both Resort hotel and City hotel booking records.

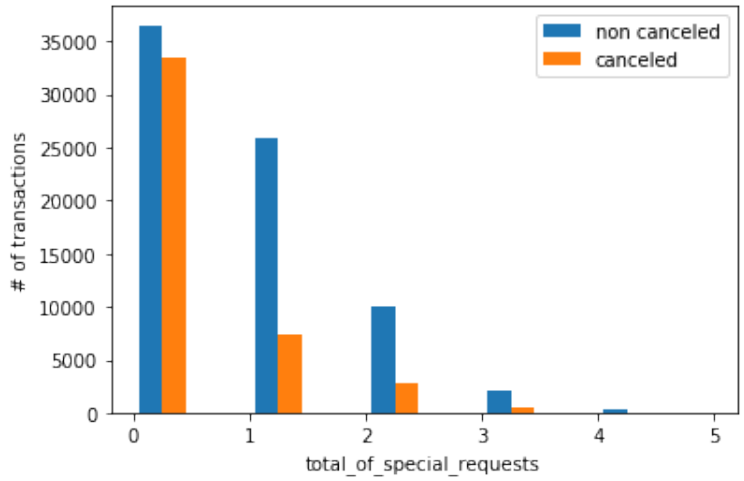



Figure 9: The # of total special requests of cancelled and non-cancelled data

### 2.2.8 Analysing the cancellation rate of each month

The graph below shows the cancel rate of each month in the year from January to December for City hotel and Resort hotel. The cancel rate is defined as the number of cancelled bookings over the number of total bookings. We can see that for the City hotel, the highest cancel rate appears in April. However, the highest cancel rate for Resort hotel appears in August and generally higher during summer. Moreover, the lowest cancel rate of City hotel appears in March while the lowest cancel rate of Resort hotel appears in January and generally lower in winter.

We found that the curve of the Resort hotel moves more gradually, while the curve of the City hotel goes up and down more frequently. Moreover, the cancel rate of the Resort hotel is generally lower than the City hotel, with the average of 27% and 41.3%, which indicates the purpose of Resort hotels are more fro planned trips and usually they are located in a more rural area so you don't have many choices to change. While City hotels are more for business travels and usually are located in a more urban area, so there might be a lot of choices for you to change to another.

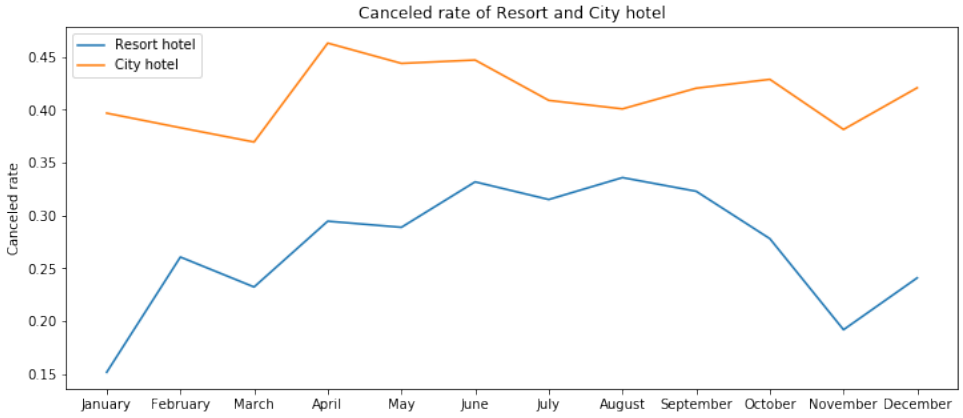



Figure 10: The cancel rate of Resort and City hotel

# 3 Cancellation Prediction

In this section, we will first construct the feature space for machine learning models based on our analysis in the last section and some specific technologies in feature engineering. Then some classic models for classification are implemented to predict cancellation status for each booking record. The comparison of the performance of different models will help us find the most suitable model for cancellation prediction in practice.

## 3.1 Feature engineering

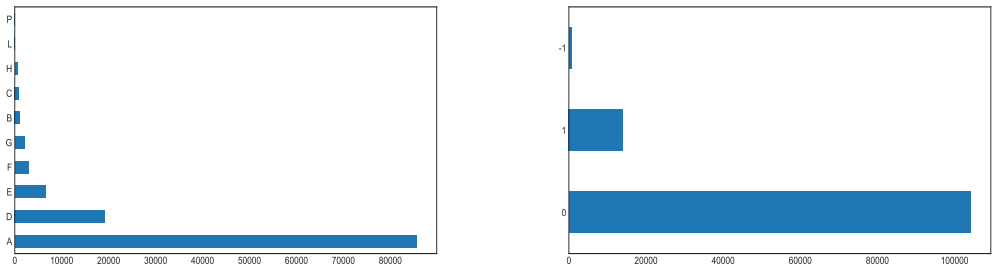
Feature engineering is crucial for machine learning algorithms since the features we use influence more than everything else the result. It is a process of using subject matter expertise to preparing a proper input data, compatible with the machine learning algorithm requirements, and thus improving the performance of machine learning models. Overall, we use three different ways to construct the feature space.

### 3.1.1 Feature combination

Feature combination is a method that combines some existing features to get some new features, which are more informative and comprehensive for the machine model to fit the dataset in a more accurate way.

- **reserved_room_type and assigned_room_type**

  For room type, what we concerned about is whether or not the guest was assigned to the same room as them booked since this might influence the decision of cancellation. Furthermore, if they did not get the room they booked, did they get an upgrade? The Figure 11a shows the distribution of the reserved room type. Since we do not have the real room type because of some business secrets, we assume an order of type to calculate the change of the room type. The new feature 'change_direct' indicates whether a customer got an upgrade (denotes by 1), a downgrade (denotes by -1), or the same room as booked (denotes by 0). The distribution of these three categories is shown in Figure 11b. In this way, the new feature has more information about the room type, which may be more correlated to the cancellation status.




(a) Old feature: Reserved_room_type (b) New feature: Change_direct

Figure 11: The features about Room_type

### 3.1.2 Feature encoding

For some category variables, we only concern about 'is' or 'is not', even though they have different types. For these variables, we convert them to binary ones.

- **agent, company**

For these two features, we only consider whether or not the room is booked through an agent or a company, and we do not focus on the information about the agent/company. Therefore, we convert these two features into binary variables where zero denotes no agent/company involved.

- **hotel**

  We convert the hotel to a binary variable as there are only two different types of hotels in our dataset: Resort hotel and City hotel. We take into account the type of hotels in the prediction since different hotels may have different cancellation rates.

For the rest categorical features, we implement the one-hot encoding method, which is commonly used in feature engineering. This method can change categorical data into the numerical format without losing any information. It spreads the values in a column to multiple flag columns and assigns 0 or 1 to them. For 'arrival_date_month', and 'arrival_date_day_of_month', we use embedding. It is a mapping of a discrete variable to a vector of continuous numbers.

### 3.1.3 Drop features

We should also drop some useless features with insufficient information since these features not only have no impact on the improvement of the model performance but also can lead to dimensional disaster.

- **country**

  For the feature country, there is a total of 177 different countries in our dataset. We group these countries according to their continent to form a more comprehensive feature to reduce the dimension. The result is shown in the Section 2.2, Figure 2. As more than 90% of the booking is from Europe, this feature is not useful in prediction. Therefore, we drop it.

- **reservation_status**

  Besides the is_cancelled feature, the dataset also includes reservation_status, which gives the last status of the reservation. The analysis of the reservation_status with is_cancelled is shown in Figure 12.

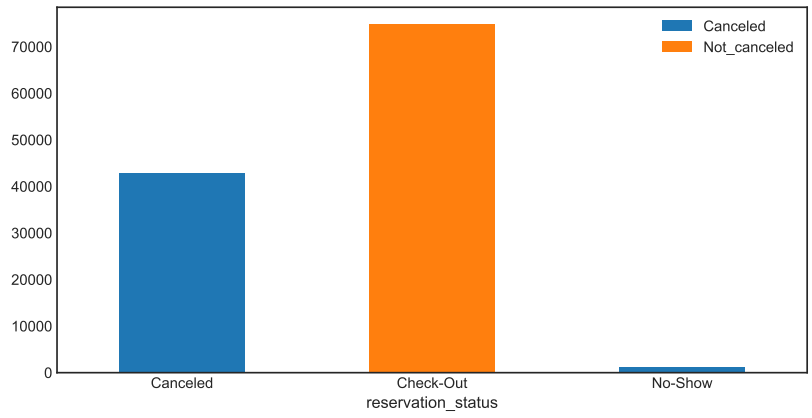



Figure 12: Relationship between is_cancelled and reservation_status

From the figure, we can see when the record is not cancelled, the status of the reservation is Check-out, and when the record is cancelled, the reservation status is either cancelled or no-show. Therefore, we drop the feature reservation_status since the feature

is_cancelled is highly correlated to the feature reservation_status. Moreover, the feature reservation_status is got after the booking is closed; thus, it is not helpful for cancellation prediction. Therefore, we drop reservation_status as well as reservation_status_date.

- **arrival_date_year, arrival_date_week_number**

  Since there was no big event that happened in 2015, 2016, and 2017, and the low season and high season of tourism are basically unchanged every year. We drop the feature arrival_date_year. Moreover, as we already have the information about arrival_date_month, we could drop the feature arrival_date_week_number.

After preparing the features, we split the dataset into train, validation, and test sets. Firstly, we split the dataset into the train set and test set. Then, we split 20 percent of the training set as a validation set. The proportion of each set is shown in Figure 13. From this figure, we can see that the proportion of cancellation in each set is almost the same, and all sets are nearly balanced, which reduces the overfitting problem in the training process. Finally, we normalize each set by standard normalization to remove the impact of the data scale.

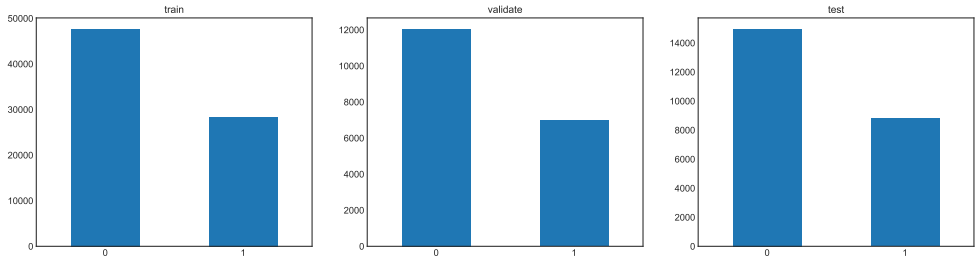



Figure 13: The proportion of cancellation in train, validation and test set

## 3.2 Supervised machine learning

In this part, our target is predicting cancellation status for each booking record based on the remaining features after the feature engineering process, which is a binary classification problem.

The data property and model complexity should be figured out before choosing the model to fit the training samples properly. Besides the model itself, the hyper-parameters in the model, which are set manually before the training process, are related to the model complexity and crucial to the model performance. A complex model often leads to overfitting to some aspects of the training samples without generality and robustness. On the contrary, a simple model can not learn all the properties of the data resulting in a low accuracy in prediction. Therefore, in order to find the optimal hyper-parameters for different models, we implement 5-fold cross validation on the training set and do grid search on the underlying hyper-parameter space. The choice of hyper-parameter space is determined by each model respectively, and the time complexity should also be considered, since one 5-fold cross validation process contains 5 training processes. So we should find the balance between the accuracy of grid search and experiment time.

After hyper-parameter tuning and training, we use different metrics to compare different classifiers to find the most suitable one for the cancellation prediction task in practice. The detailed experiment setting and evaluation results are as follows.

### 3.2.1 Baseline models

- **Logistic regression**

  Logistic regression is a statistical linear model that involves a logistic function to model a binary dependent variable. It is one of the most common machine learning method used to solve the binary classification (0 or 1) problem based on a similar idea as regression and to estimate the possibility of each binary label.

  In our experiment, we implement logistic regression with learned feature augmentation and logistic loss to predict cancellation status. The loss function of logistic regression is the log loss function, which is defined as follows:

$$LogLoss = \sum_{(x,y)\in D} -ylog(y') - (1-y)log(1-y') \tag{1}$$

  where $D$ is a data set containing labeled samples $(x, y)$, $y$ is the label of the sample, and $y'$ is the predicted value.

  There is no hyper-parameter in regular logistic regression model, but it involves gradient descent to make the parameters converge to optimal ones by minimizing the value of loss function. The training process is shown in Figure 17 in the appendix. All evaluation metrics are calculated on test set.

- **Decision tree**

  Decision tree is another classic predictive model in machine learning, which is more interpretive. It represents a non-linear mapping relationship between object attributes and object values. Each node in the tree represents an object, each forked path represents a possible attribute value, and each leaf node corresponds to the value of the object represented by the path from the root node to the leaf node. The purpose of splitting attributes is to achieve the maximum information gain which is indicated by 'entropy' or 'Gini impurity'.

  Compared to logistic regression model, the decision tree is more suitable for the data containing many categorical features and do not run gradient descent to find the optimal parameters. Moreover, the result of decision tree is more understandable, since the attributes denoted by the top tier nodes are more related to the prediction label. However, decision tree is easy to overfitting as the depth of the tree increases, so we should choose the hyper-parameter more carefully.

  In the cross validation, the depth between 1 and 30 is the grid search space we set. And we choose entropy as the split criterion and find that the F1 score is the highest when the depth of the decision tree is 24 which is shown in Figure 18 in the appendix. Based on this tuned hyper-parameter, we build the decision tree to predict cancellation status and compare it to the ensemble models introduced in the next part.

#### 3.2.2 Ensemble models

In the supervised machine learning, our goal is to learn a stable model that performs well in all aspects, but the actual situation is often not so ideal, and sometimes we can only get multiple models with preferences (the weakly supervised model performs better in some respects). Ensemble learning can combine multiple weakly supervised models in order to get a more comprehensive robust supervised model. The underlying idea of ensemble learning is that even if a weak classifier gets a wrong prediction, other weak classifiers can correct the error. So we implement two ensemble models to find a more accurate classifier.

- **Random forest**

  Random forest is a kind of bagging ensemble model, which is consisted of several independent decision trees, and the output prediction is based on the aggregation of the prediction of each decision tree. Bootstrap sampling is implemented to generate several sampled training dataset with the same size as the original training dataset. In this way, the bias between positive and negative samples can be smoothed and the generality and anti-disturbance property increase. So the random forest has a better performance on the imbalanced dataset and the dataset with noise and outliers.

  In the random forest model, the crucial hyper-parameters are the number of trees, the depth of each tree, and the maximum feature of each tree. In the cross validation, we try different numbers of trees from 10 to 150, but the F1 score changes slightly, so we choose 100 as the number of trees. For the depth, we find that 24 is the best depth of the decision tree, so we keep the same depth in the random forest. For the maximum feature, we choose 18 as the maximum feature which corresponds to the highest F1 score on the validation set which is shown in Figure 21 in the appendix.

- **AdaBoost**

  AdaBoost is the abbreviation of 'Adaptive Boosting' and the typical boosting ensemble model. The self-adaptation lies in that the samples divided by the previous basic classifier will be strengthened, and the weighted samples will be used again to train the next basic classifier. At the same time, a new weak classifier is added in each round until it reaches a predetermined sufficiently small error rate or the maximum number of iterations specified in advance. AdaBoost can be implemented in most of classification tasks and the modification of parameters is easy. But it is sensitive to the outliers in training data.

  In our experiment, we use decision tree as the basic classifier in AdaBoost. The learning rate should be chosen in AdaBoost and we choose 1 as the learning rate based on the results of the cross validation shown in Figure 23 in the appendix.

### 3.2.3 Model evaluation

ROC curve is a common method to evaluate the classifier. It is indicated by false positive rate and true positive rate. The classifier with the highest area under the ROC curve (AUC) has the best performance. In Figure 14, we can find that the random forest classifier has the highest AUC score (0.90) and the decision tree classifier has the lowest AUC score (0.84).

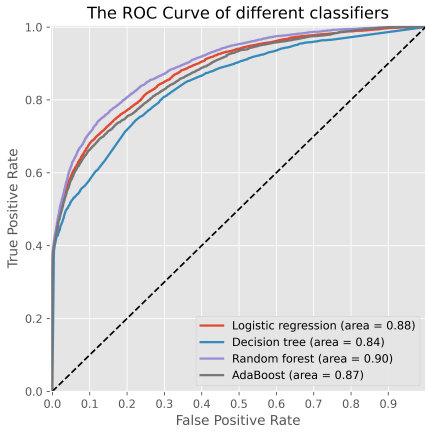
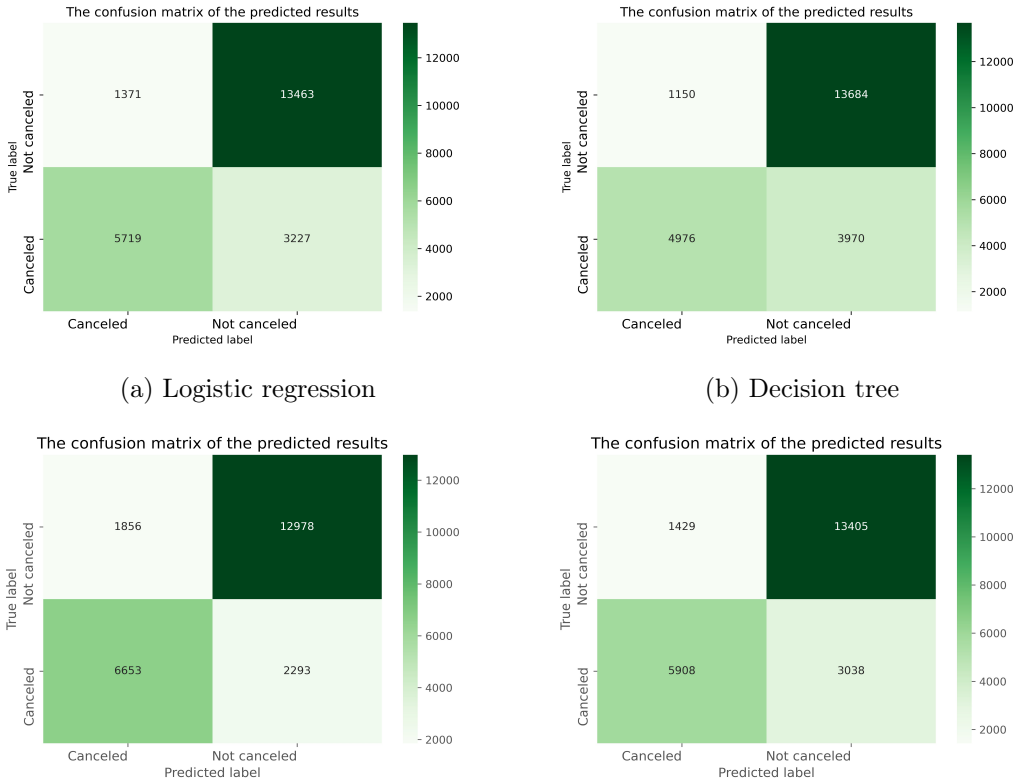



Figure 14: ROC curves of different classifiers

When we evaluate some particular classifier, we should not only consider the accuracy but also the precision and recall, since the false positive and true negative samples affect the overall performance severely in practice. So we use the F1 score to compare different classifiers based on confusion matrices in Figure 15. We can find that the random forest classifier has the best F1 score (0.76) and the decision tree has the lowest F1 score (0.66).

Moreover, in practice, if a cancelled record can not be detected (corresponding to the true negative in the confusion matrix), it will affect the booking system of the hotel and cause economic loss. So we want to find the classifier with the lowest true negative rate. And the random forest classifier also has the minimum true negative samples.




(a) Logistic regression

(b) Decision tree

(c) Random forest

(d) AdaBoost

Figure 15: Confusion matrices of different classifiers

Therefore, the random forest classifier has the overall best performance. And we will use the feature importance generated by random forest classifier to find some business ideas in the following part.

# 4 Business View

## 4.1 Discover Demand, Create Demand

From our EDA we observed that a majority of the hotel customers who cancelled had no request. Our analysis of cancellation and requests is interesting, although the number of cancelled and non-cancelled both decrease with the increase of request, in general, the proportion of zero requests for non-cancelled is greater than cancelled. The logic behind should be well understood, if there is a request, there is a demand; if there is more demand, it means the hotel matches well. However, this is a tricky problem, because hotels, at least in this dataset, were not aware of the value behind those random/potential customers. Customers who do not request can be transformed and retained.

Five-star hotels serve different customers from budget hotel chains. Budget hotel chains, which provide basic accommodation, low prices, fast service to attract general business guests and self-service tourists; hotels such as Hilton and Shangri-la are symbols of comfort, luxury and status. When desires and purchasing power of customers are fixed, they choose products that best meet their needs. Therefore, the task of the hotel is to determine the demand of target customers through a series of market research, and then provide customers with the greatest satisfaction of the product.

The demand has characteristics such as diversity, developmental and gradation, it will change with social and technological progress and economic development. This will involve a deep-seated problem, that is, the customer does not know whether there is a demand, namely potential demand, which depends on the hotel to create, dig and induce. For example, since ancient times, people in China did not have the habit of drinking coffee. After coffee was packaged and promoted as a symbol of tasteful life, young people regarded coffee as a romance in the west, and coffee culture started its conquest in China. It can be said that the demand was created.

Online booking accounts for a large part of the product promotion in the hotel industry, which requires the hotel front desk staff to explore potential demands as much as possible within a short period of contact with customers. The front desk staff need to first understand the needs of customers, introduces and publicizes the product they need, then get the information of customers, and deeply explore the potential demands, and finally guide the demands of customers to their advantages. Customers are not necessarily able to make the right choice on how to realize their potential demands, which requires our sales staff to first explore the causes of the problem, find out the internal needs through the external appearances, and then guide the customer demands to their products.

## 4.2 Incremental Competition, Stock Competition

Our attempts to look at user behaviour from an intercontinental perspective, away from the state, have yielded a startling result: 90% of the bookings were from Europe. This might make hotel operators restless. The hotels should not rely only on European visitors and should explore strategies to attract more visitors from other continents such as Asia, South America and Africa, where they already have some from countries outside of Europe.

The key contradiction we have to solve is the contradiction between consumers' demand for accommodation experience and the imbalance and insufficiency of the hotel industry. At present, the demand of consumers is increasingly diversified, and the formats and products of housing supply are becoming more and more diverse. However, there is still an imbalance in the regional level structure of supply and inadequate supply types and contents, which require us to study the needs of consumers carefully, use big data to portray consumers, truly make people-oriented, and match the accommodation supply according to the needs of consumers. It seems simple, but

very difficult.

But we should also be aware that the emergence of an incremental market is a very rare phenomenon for mature hotel industry. The future of the hotel industry must be from the explosive incremental market gradually to the existing stock market based on the market background. The growth will tend to be small, only in the original market to find and there will be industry peaks, bankruptcies, mergers and acquisitions. At the same time, it will be accompanied by continuous innovation and industrial upgrading.

This raises a new question, facing the untapped non-European market, should the hotel actively participate in the competition? In our opinion, the incremental market is not an easy path either. Many new market segments are limited by the environment, and the outbreak of Covid-19 cast a shadow on the future, and might cause the decline and downturn of incremental market demand. Therefore, in the complex and changing market environment, for many hotels, it is necessary to readjust the means of market expansion and marketing mode. No matter in stock or incremental market, we are faced with many challenges and pressures. Only by breaking the boundary of 'stock and increment' and truly innovating service products, can we find all opportunities and spaces that can be utilized and seize the customer.

## 4.3 Feature Importance

Through machine learning, we can use feature importance to identify the features that most relevant to hotel cancellation prediction, which provides a theoretical basis for the strategic direction of the hotel. In our study, we found that the top 5 important features are 'deposit_type_Non Refund', 'deposit_type_No Deposit', 'lead_time', 'total_of_special_requests' and 'adr'.

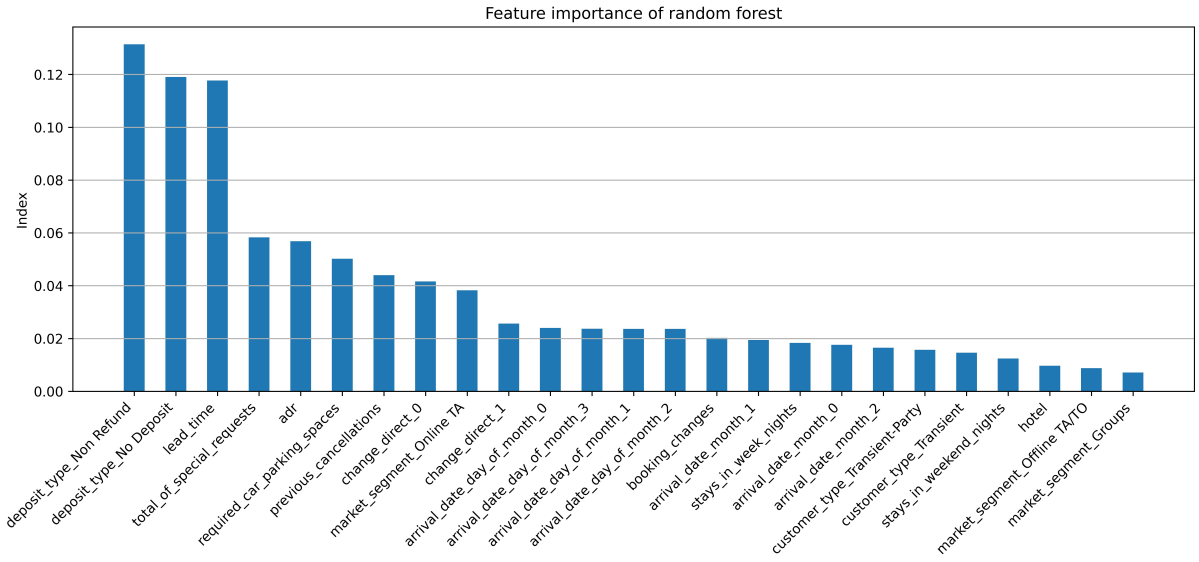



Figure 16: Feature importance of random forest

### 4.3.1 Analysis about deposit

We found deposit is highly related to final cancellation status and this is consistent with consumer behaviour. 'deposit_type_Non Refund' will probably be cancelled less, but 'deposit_type_No Deposit' will be cancelled more. We need to study the psychology of customers. A rational judgment is that we should set a certain amount of deposit, which should not be so high as to scare off consumers; it also cannot be too low, because it does not have the binding force.

Customer manager should also have the consciousness of classification, for different types of room in a hotel, the setting of the deposit should also be diverse, because consumers are price sensitive.

### 4.3.2 Analysis about lead_time

Lead times can help reduce hotel cancellations. Today, markets are dynamic and uncertainty is the new normal. Hence, a short lead time is not automatically considered bad, after all, last-minute bookings are usually very profitable. Hotels can consider changing the cancellation conditions by setting a maximum advance booking limit. Therefore, you can reduce the lead times and the number of cancellations.

However, it is important to note that setting such restrictions may negatively impact the sales and bottom line. Longer lead times are preferable to hotels because they offer many benefits, few finance directors would reject cash flow benefits or the ability to anticipate staffing schedules more cost-effectively. And from a revenue management perspective, it gives hotels more time to move inventory to direct channels, and reconsider other deals, to maximize revenue.

### 4.3.3 Analysis about adr

ADR is Average Daily Rate, We divide actual daily room revenue by total rooms sold to get it. It is one of the most important indicators for hoteliers to measure hotel performance. The idea of optimizing ADR can enable us to resist the risks brought by cancellation, or reduce the cancellation rate.

A good way to improve 'adr' is promotion. Just as guests shop around for the best price, they also look for attractive promotions. When you offer a promotion, you'll have an edge on the competition and naturally make your hotel look more attractive. Not only will people take advantage of the promotion, but they will probably tell their friends and family about it, which makes a big splash in your hotel.

# 5 Conclusion

In this project, we analyze the Hotel Bookings dataset and try to find the most suitable machine learning model to predict whether a booking will be cancelled, which is important to the booking system and revenue of a hotel.

In the exploratory data analysis, we can find some interesting differences between cancelled and non-cancelled hotel booking records without implementing machine learning models. Firstly, we find that in the cancelled bookings, a lot of them are booked from an earlier date. Secondly, we find that the customer's previous cancellation history is relevant to the current booking. If the customer has cancelled his booking at least once before, he is more likely to cancel his current booking. However, if this customer has a lot of booking records that are not cancelled, it is more likely that he is also going to keep this current booking. Moreover, by analyzing the customer's booking changes and special requests, we find that for the bookings that are not cancelled, they generally have more booking changes and special requests. This could be caused by that they care more about the trip and wish a perfect trip so they make changes and requests to ensure the quality of their trip and is less likely to cancel the bookings. By analyzing the customer's market segment and the number of people in the booking orders, we find that group customers with several adults are more likely to cancel their bookings while customers who come to the hotel directly to make the booking are less likely to cancel their bookings. Finally, when looking into the cancellation rate of each month in the year of Resort hotel and City hotel, the cancellation rate of City hotel is strictly higher than that of Resort hotel, which indicates the difference of the original purpose and location of these two kinds of hotels. Their generic marketing strategy and target markets are also different.

For the machine learning part, we find that the ensemble models overall perform better than baseline models. However, there are also more hyper-parameters that should be chosen in ensemble models. When the amount of bookings is high, it is very time-consuming in hyper-parameter tuning. So we should find a balance between model complexity and performance. Regardless of the model complexity, the random forest model has the best F1 score and AUC score, which indicates that it is the most suitable classifier. Another crucial factor we should consider is the true negative samples. When a cancelled booking could not be detected by the classifier, it will affect the room pricing and room inventory significantly. And the random forest classifier can also achieve the lowest true negative rate.

With an investment of 400 million five-star hotels, it is estimated that the overall decoration of the hotel in the hospitality industry will be carried out after 10 years, and 50% of the equipment will be replaced, the depreciation cost of decoration will be 20 million every year. The hotel industry is a low labour efficiency per capita industry, but also a slow-changing industry. Nevertheless, there is few data analysis regarding customer services and few data transmission between industrial chains.

The problem faced by the hotel is not to collect data, but to collect data and integrate them, and put them into practice. As a result, there is no doubt that hotels (and we) face challenges in determining how and where to start the data analysis process, and these challenges even hinder many people from trying to do so. Many hotels do not make full use of the relationship they have or should have with their customers, make good use of this relationship, and build a better system based on these data, which can become a competitive advantage.

As you can see, our analysis is a good example. Under the data-driven machine learning tool, we explore the opportunities of hotels, such as potential market and user demand. But we should also note that as new business models continue to emerge, old forecasting methods that are effective in moving forward may not work. Researchers need to continue to develop new and better predictions. However, the prediction model can not be applied to all hotels, so hotels should also

actively explore their model. With the rapid development of machine learning, researches and advanced models in the hotel industry will soon lead to a competition for business transformation capabilities. It's time for hotel decision-makers to be prepared to take advantage of this huge, and probably unprecedented opportunity.

# A Appendix

## A.1 Figures in logistic regression

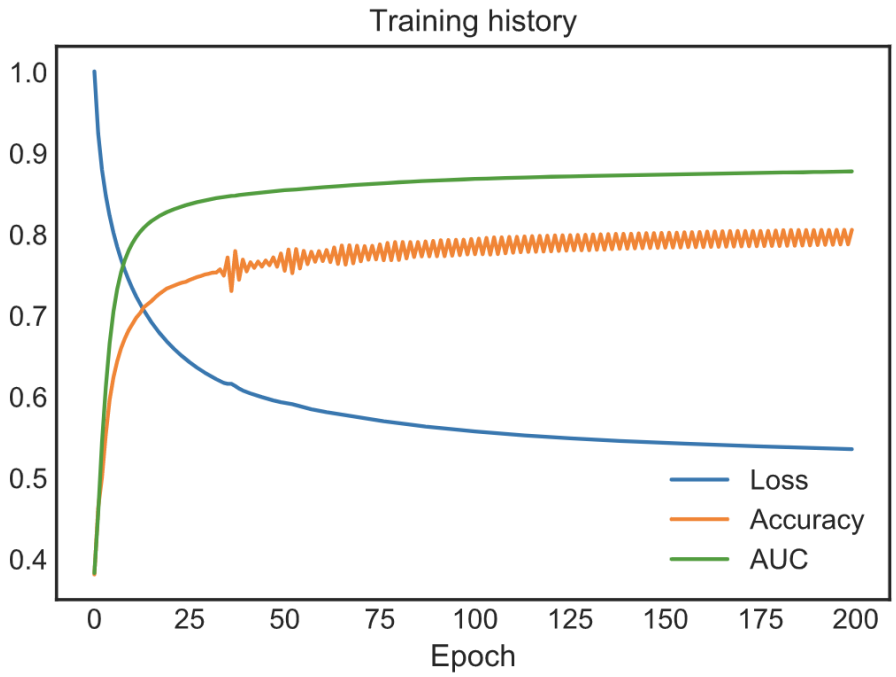



Figure 17: Training history of logistic regression

In Figure 17, we can see that the algorithm converges after 50 epochs approximately. The accuracy on test set reaches around 0.78 and AUC score reaches around 0.87.

## A.2 Figures in decision tree

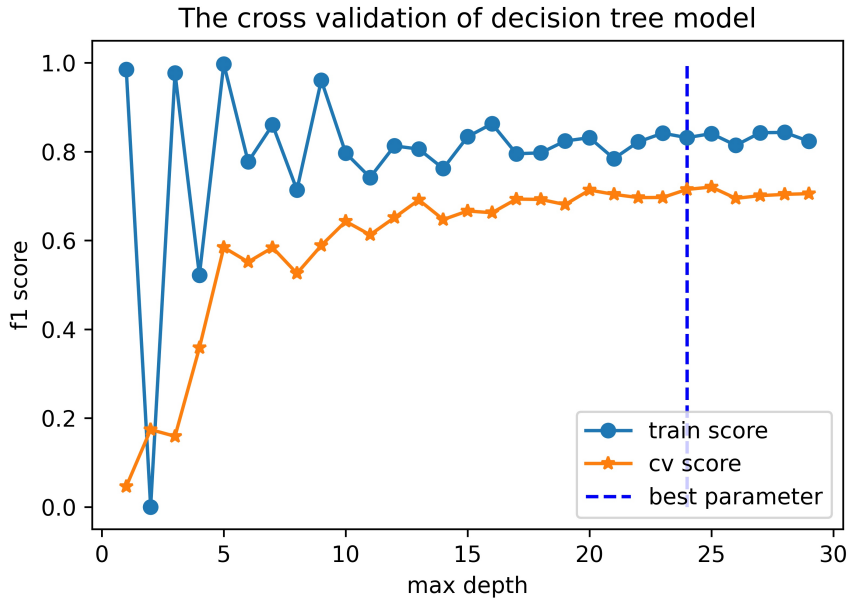



Figure 18: Cross validation of decision tree (grid search on max depth)

In Figure 18, the train score is calculated on the training set directly and the cv score is calculated by cross validation process on the training set. We can see that when the max depth equals to 24, the cv score is the highest.

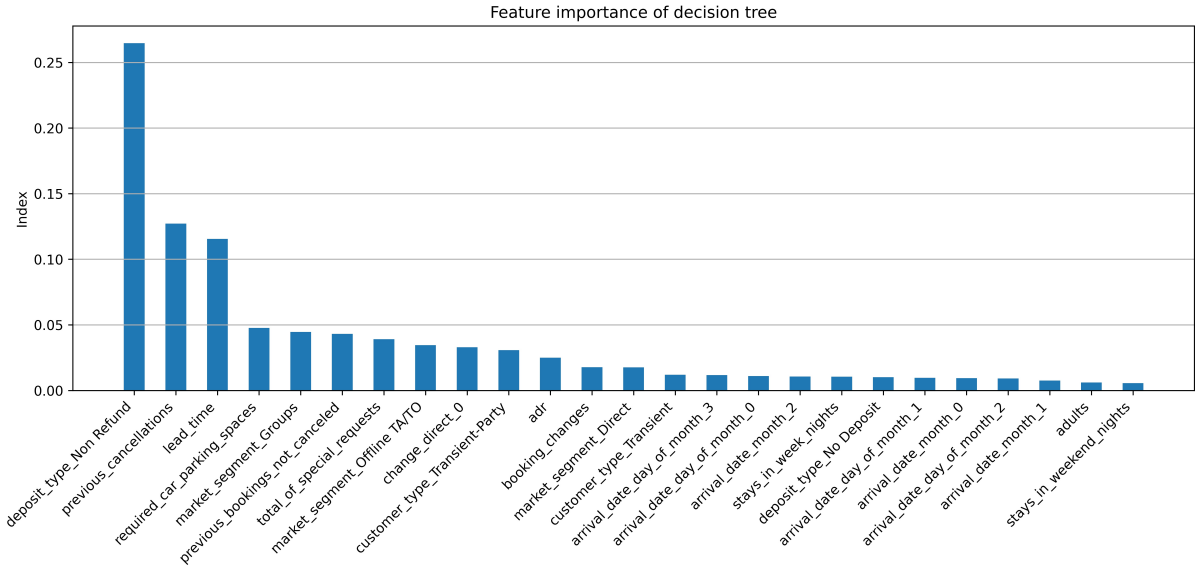



Figure 19: The feature importance of decision tree

In Figure 19, we only keep the top 25 important features of decision tree due to space limitation. 'Deposit_type_Non Refund' is the most important feature with an index around 0.25. 'Previous_cancellations' and 'lead_time' are also closely related to the cancellation status.

### A.3 Figures in random forest

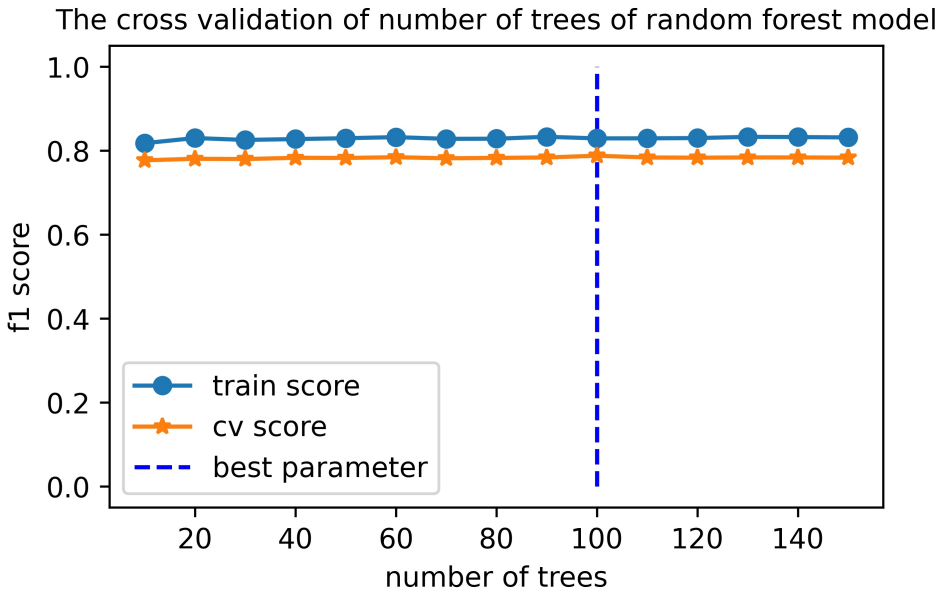



Figure 20: Cross validation of random forest (grid search on the number of trees)

In Figure 20, we can find that the number of trees does not affect the F1 score obviously. So we choose 100 as the number of trees and do grid search on another hyper-parameter.

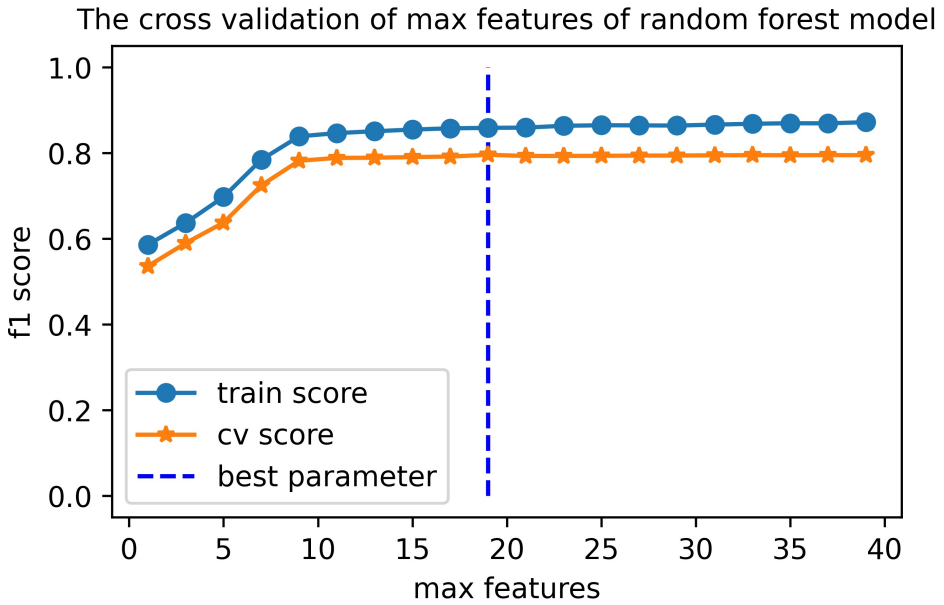



Figure 21: Cross validation of random forest (grid search on max feature)

In Figure 21, we can find that the F1 score increases as the max feature increasing. So we set 18 as the max feature which performs best in cross validation.

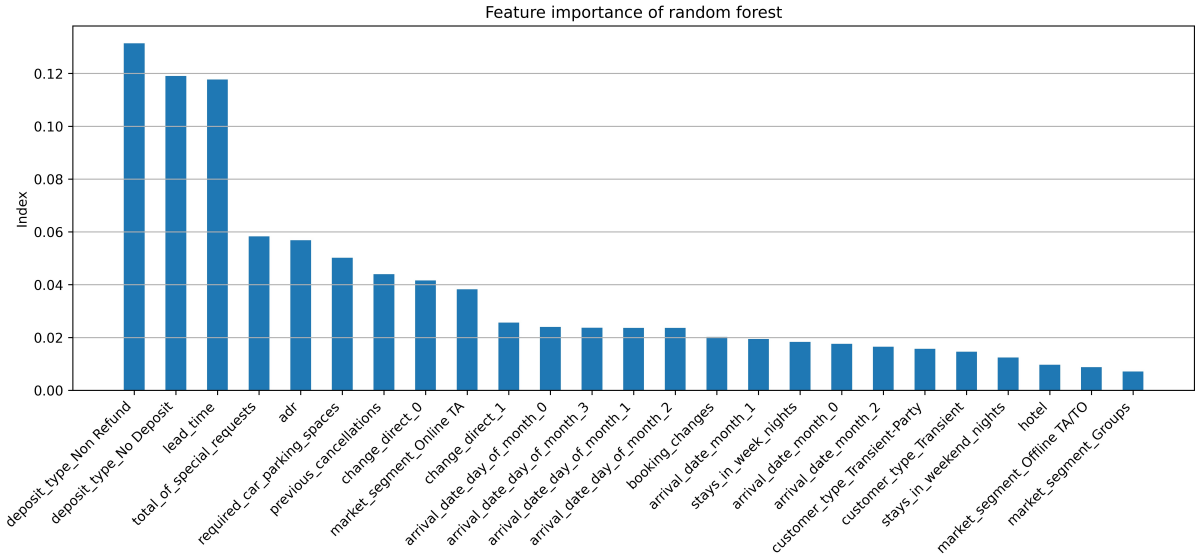



Figure 22: The feature importance of random forest

In Figure 22, we can find that the indexes of each feature are more uniform than those of the decision tree. The 'deposit type', 'lead_time' and 'total_of_special_requests' are the most important features.

## A.4 Figures in AdaBoost

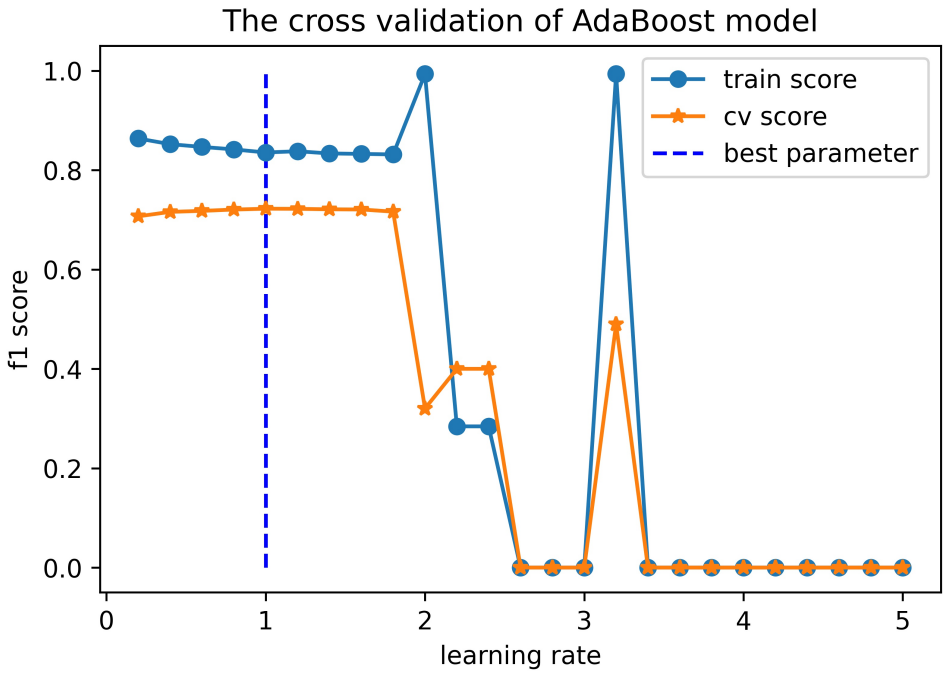



Figure 23: Cross validation of AdaBoost (grid search on learning rate)

In Figure 23, we can find that when the learning rate is more than 2, the F1 score decreases dramatically. And we choose 1 as the learning rate.

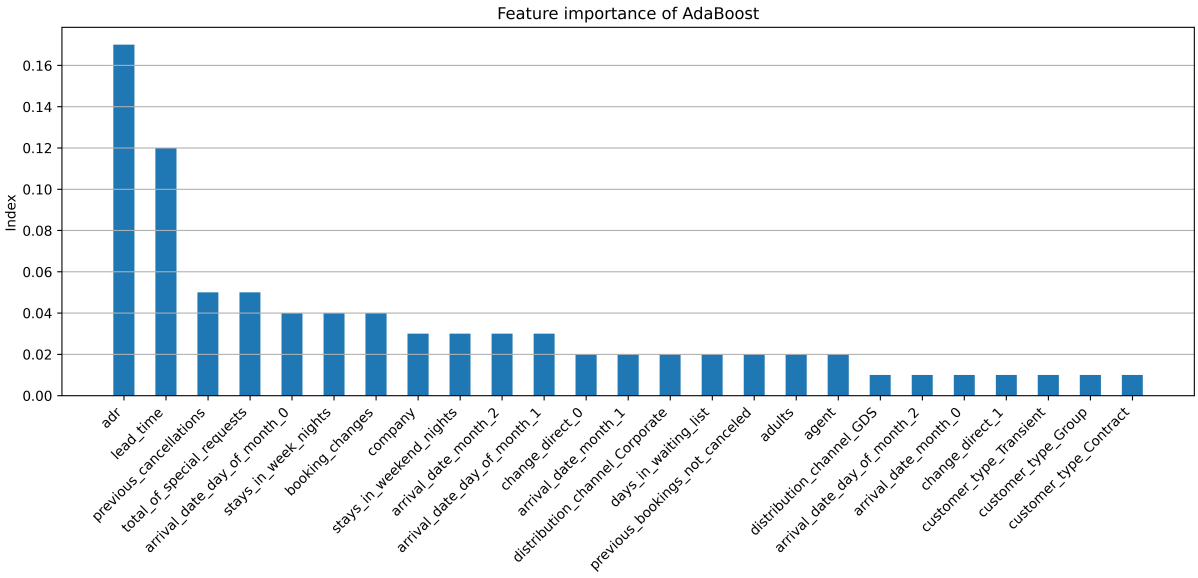



Figure 24: The feature importance of AdaBoost

In Figure 24, the most important feature is 'adr', which is different from the previous results. However, 'lead_time' and 'previous_cancellations' are also closely related to the cancellation status of a booking.

## References


American Hotel & Lodging Association, Uniform System of Accounts for the Lodging Industry, 11th Revised edition,.Educational Institute, New York, 2014.

Antonio, Almeida, Nunes. *Predicting hotel bookings cancellation with a machine learning classification model*, in: Proceedings of the 16th IEEE International Conference Machine Learning Application, IEEE, Cancun, Mexicopp. 1049–1054. doi:10.1109/ICMLA.2017.00-11, 2017.

Antonio, Almeida, Nunes. *Hotel booking demand datasets.* Published in Elsevier Inc, doi: 10.1016/j.dib.2018.11.126

Chen, C.-C., Schwartz, Z., & Vargas, P. (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), 129–135.

Chen, C.-C., & Xie, K. (Lijia). (2013).*Differentiation of cancellation policies in the U.S. hotel industry.* International Journal of Hospitality Management, 34, 66–72.

Chiang, W.-C., Chen, J. C., & Xu, X. (2007). *An overview of research on revenue management: current issues and future research.* International Journal of Revenue Management, 1(1), 97–128.

De Lollis, B., 2002. *Hotels get less hospitable about the late cancellations.* USA Today, August 27, 6B

DeKay, F., Yates, B., Toh, R.S., 2004. *Non-performance penalties in the hotel industry.* International Journal of Hospitality Management 23 (3),

Engle, J., 2009. *Hotel cancellations can cost you.* `http://articles.latimes.com/2009/oct/25/travel/tr-money25`.

Koide, T., Ishii, H., 2005. *The hotel yield management with two types of room prices, overbooking and cancellations.*. International Journal of Production Economics 93–94, 417–428.

Lindenmeier, J., Tscheulin, D., 2008. *The effects of inventory control and denied boarding on customer satisfaction: the case of capacity-based airline revenue management.* Tourism Management 29 (1), 32–43.

Mandelbaum, R. (2016). *How attrition, cancellation fees hit your bottom line.* hotelnewsnow.com website at `http://www.hotelnewsnow.com/Articles/54143.`

Mehrotra, R., & Ruttley, J. (2006). *Revenue management (second ed.).* Washington, DC, USA: American Hotel & Lodging Association (AHLA).

Moe, W., Fader, P.S., 2002. *Using advance purchase orders to forecast new product sales.* Marketing Science 21 (3), 347–364.

Morales, D. R., & Wang, J. (2010). *Forecasting cancellation rates for services booking revenue management using data mining.* European Journal of Operational Research, 202(2), 554–562.

Phillips, R., 2005. *Pricing and Revenue Optimization.* Stanford Business Books. Stanford University Press, Stanford, California.

Schwartz, Z., 2000. *Changes in hotel guests' willingness to pay as the date of stay draws closer.* Journal of Hospitality & Tourism Research 24 (2), 180–198.

Schwartz, Z., 2006. *Advanced booking and revenue management: room rates and the consumers' strategic zones.* International Journal of Hospitality Management 25 (3), 447–462.

Schwartz, Z., 2008. *Time, price and advanced booking of hotel rooms.* International Journal of Hospitality and Tourism Administration 9 (2), 128–146.

Shugan, S.M., Xie, J., 2005. *Advance-selling as a competitive marketing tool.* International Journal of Research in Marketing 22 (3), 351–373.

Talluri, K. T., & Van Ryzin, G. (2004). *The theory and practice of revenue management.* Boston, MA, USA: Kluwer Academic Publishers.

Takizawa, N. (2017). *Realities of hotel reservation, a big problem of cancellations at the last minute (in Japanese).* bunshun.jp website at `http://bunshun.jp/articles/-/2370`.

Wangenheim, F., Bayón, T., 2007. *Behavioral consequences of overbooking service capacity.* Journal of Marketing 71 (4), 36–47

Xie, J., Gerstner, E., 2007. *Service escape: profiting from customer cancelations.*Marketing Science 26 (1),