Data Mining for Business

Hotel Profits and Management

Can we predict whether a customer will cancel the hotel reservation?

By Team:

Nicolas Clarke de Dromantin

Jifeng Li

Weifan Lin

Yinzhe Lu

Rasune A Trazie

Timothy Welles

December 2020

Data Mining For Business

Instructor: Dr. Lin Hao

Table of Contents

Ⅰ. Executive Summary

The hotel industry is one of the most competitive industries around the world. It is also an industry that we are all familiar with and have a strong connection with, making this project more interesting and engaging. We notice that hotel operations suffer losses when facing unexpected cancellations. Cancelations are remarkably costly for hotels and strongly impact their yearly revenues.

Our project predicts whether or not a customer will cancel their hotel-booking, which determines whether the hotel should arrange resources to prepare the room for the booking and implement appropriate measures, e.g., overbooking the room. Our team decides to address this problem by predicting whether the hotel-booking will be canceled. The main advantage that hotels will gain from our predictions is that they can better decide if they shall arrange resources to prepare the room for reservation, eventually minimize the cost and maximize the profits. To achieve our goals, we followed the necessary steps. We also performed some additional exploratory analysis and variable transformation in order to clean up the data. We performed numerous data mining techniques using Python in order to extract the most important and relevant predictors. We ran Logistic Regression, Random Forest and Classification Tree models on this large dataset. Our results, charts, graphs, analysis, predictions, conclusions, and recommendations are presented in this report.

Ⅱ. Background Information

Our dataset is a Kaggle dataset named “Hotel Booking Demand” published in February 2019. This dataset features 119,390 individual entries and 32 variable columns during 2015-2017. The data is structured, and the variables are a mix of categorical and numerical types. We chose this dataset because we see the potential to bring a real business advantage to the hotel industry. Besides, all of our team members can relate to booking and canceling a hotel room. By analyzing the data, we are looking for insights that can predict the cancellation results.

The dependent variable is whether or not a customer will cancel the hotel reservation. We have built a DIDA framework to analyze this dataset to move forward with our prediction model.

Data: “Hotel Booking Demand” from Kaggle
Insights: Probability of canceling a hotel reservation
Decision:

- Whether or not it is likely to receive a hotel-cancelation from a customer?
- Whether or not to prepare the room for reservation?

Advantage: Avoid putting resources in preparing the room for potential canceled reservation; decrease the cost and eventually increase the profits; forecast sales and net profits;

Our type of Insight is a probability based on our dependent variable. Hotel cancelation is a set of binaries. The dataset is an individual-level data meaning that each row represents a single transaction. We will analyze historical data with binary dependent variables, including both cases (‘is_canceled’ and ‘is_not_canceled’) of each past transaction to predict a probability.

Based on our institutional knowledge, we have relevant predictors which include: resort hotel or city hotel, lead time, arrival date, stay-in nights, weekend or weekday, country of origin, with or without children and babies, repeated guest, cost of the reservation, room type, customer type. As each transaction has happened already, the predictors are exante.

As the number of observations of this dataset is larger than the minimum required number of observations (n=396), the dataset satisfies the portrait-shape requirement. Calculation of the minimum required number of observations:

m: number of classes in yes/no dependent variable; u: number of predators; n: minimum number of observations:

$$n = 6 * m * (u+1) = 6 * 2 * (32 + 1) = 396$$

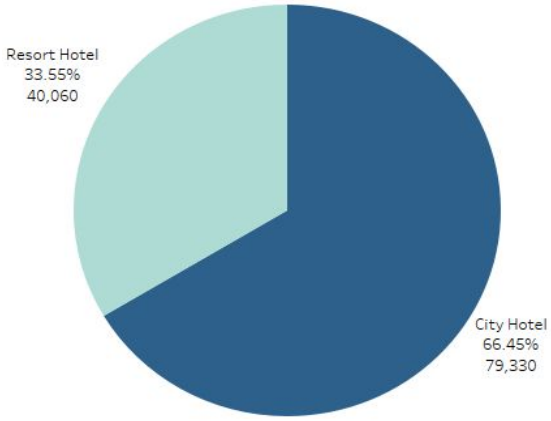
Thus, the number of data entries is greater than what a sufficient prediction model requires.

Ⅲ. Exploratory Analysis and Variable Transformation

In this step, we performed data visualization to overview the composition of the variables and the distribution of different components. Now we can get familiar with the circumstances of variables, which enables us to execute proper variable transformation for some variables.

1). Hotel type

Regarding categorical variable 'hotel', it is composed of two categories, Resort hotel and City hotel. According to the graph, Resort hotel occupies 33.55% while City hotel occupies 66.45% among all the hotel bookings.

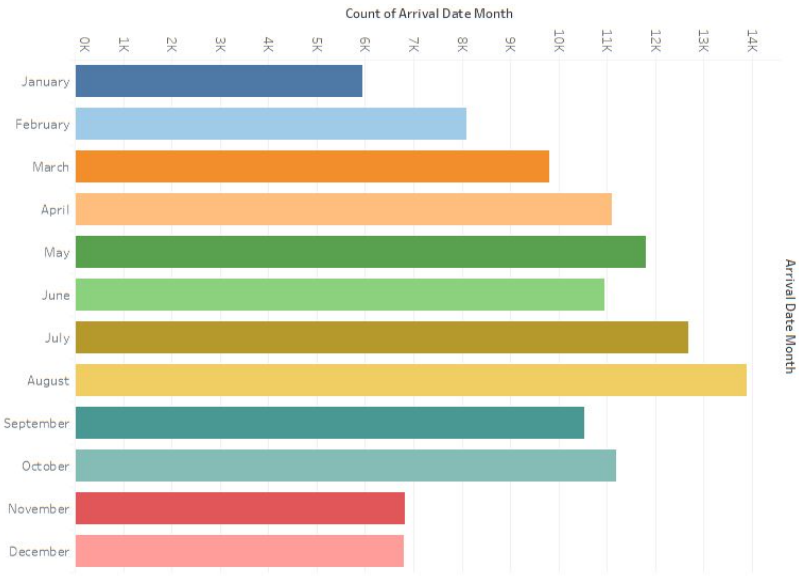


2). Arrival of Month

Regarding, categorical variable 'Arrival date of Month', it is composed of 12 months. From the graph, we can see that most people are booking hotels in the summertime, May to August. Detecting this characteristic, we learn from the Portugal Travel and Tourism Administration that June, July, and August are the peak season.

So eventually, we transformed the variable 'arrival data of month' into

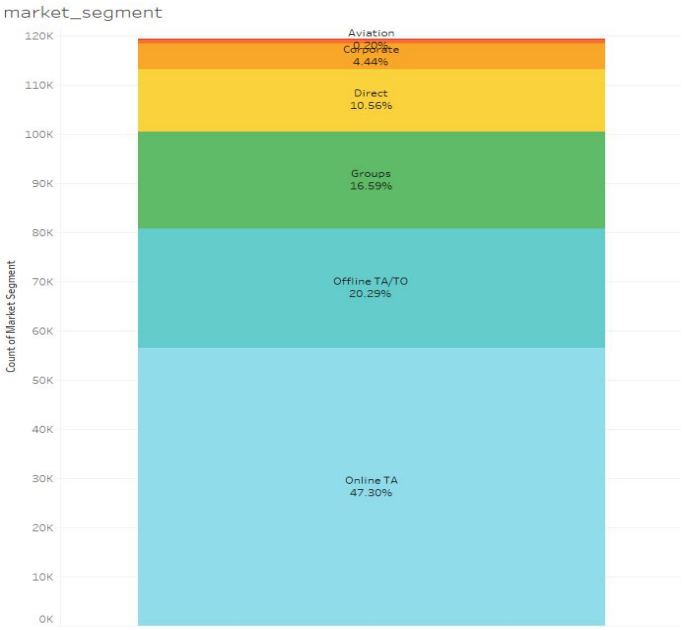
- 'peak_season', consisting of June, July and Aug;
- 'winter_season', consisting of November, December, January and February;
- 'middle_season', consisting of March, April, May, September, and October.



3). Market Segment

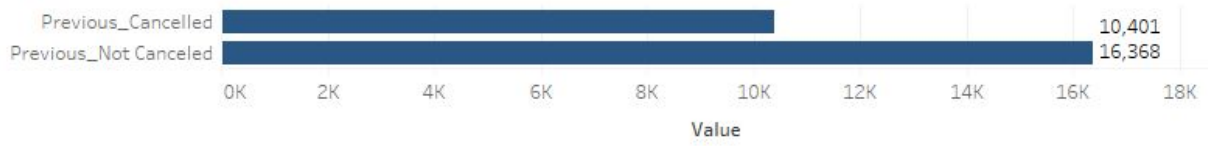
Regarding the categorical variable 'Market segment', it is composed of 7 categories. According to the graph, nearly 50% (47.3%) of hotel-booking are made through Online Travel Agency and nearly 20% of hotel-booking are made through Offline Travel Agency. 67.3% of consumers are going to an agency, online or offline for hotel-reservation. Additionally, 16.69% of the bookings are made by Groups and 10.56% of the bookings are made by consumers themselves directly.

As Offline Travel Agency is the majority market segment, we transformed the variable 'Market Segment' into 'online_purchased' and 'not_online_purchased'.



4). Previous cancelation vs. Previous not canceled

When we looked at two variables, 'Previous_cancelation' and 'Previous_bookings_not-canceled', we found that more than 61% of hotel-bookings belong to the consumers who have no previous cancelation while 38.85% of hotel-bookings belong to those who have previous cancelations.

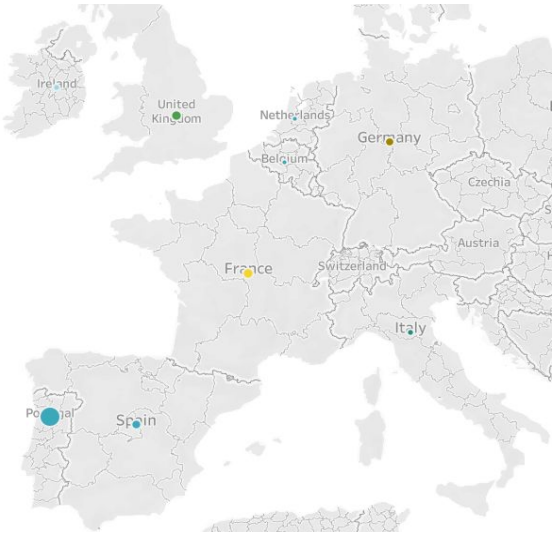


5). Country

Regarding the categorical variable 'Country', it is composed of 178 countries in total.

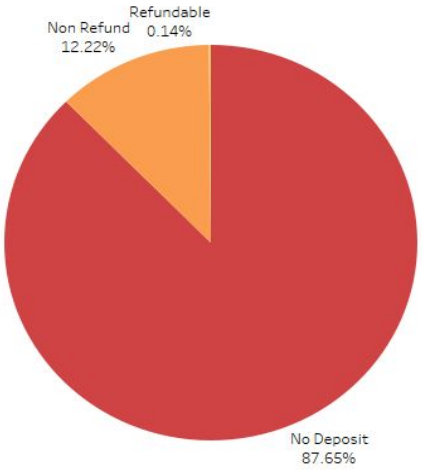
We found the TOP 10 countries are mostly European. They are Portugal, UK(GBR), France, Spain(ESP), Germany(DEU), Italy, Ireland, Belgium(BEL), Brazil(BRA), and the Netherlands.

We also found that 40.87% of the consumers are from Portugal. Realizing that this could be a potentially highly skewed distribution, we transformed the variable 'Country' into 'origin_Portugal' and 'orgin_other_countries'.



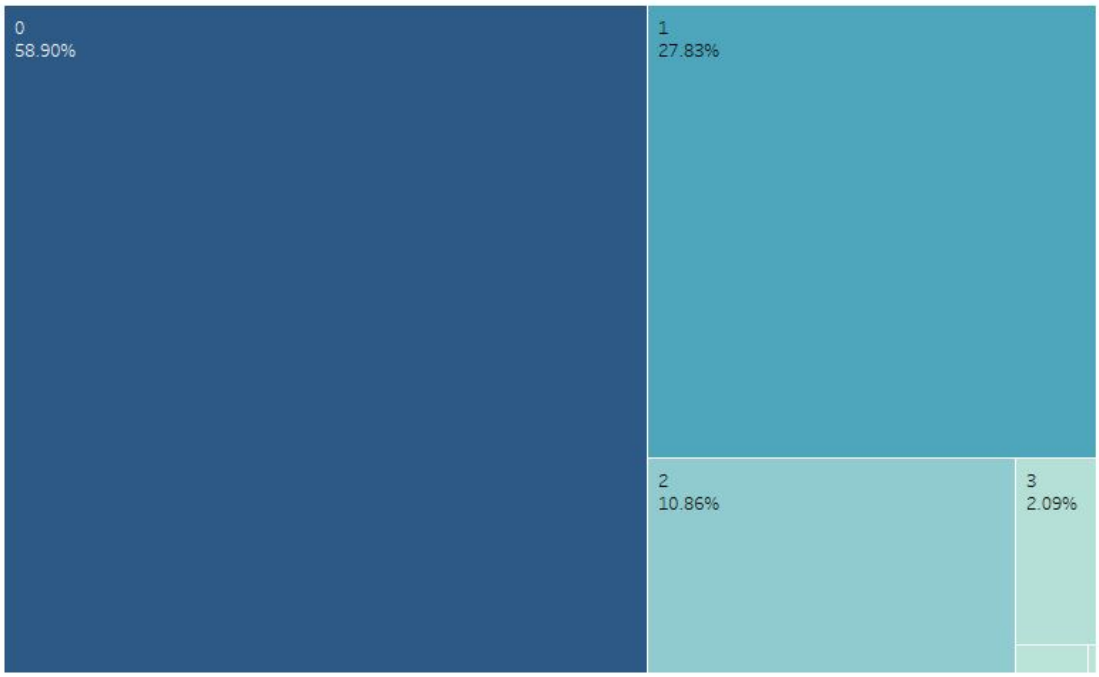
6). Deposit type

Regarding the categorical variable 'Deposit Type', it is composed of 2 main categories. 87.65% of customers booked the hotel by No_Deposit payment method. 12.22% of the hotel-bookings are not refundable. Two main deposit types occupied the largest share, and no_deposit is the primary category. Hence we transformed the variable 'Deposit Type' into 'no_deposit'.



7). Special Requests

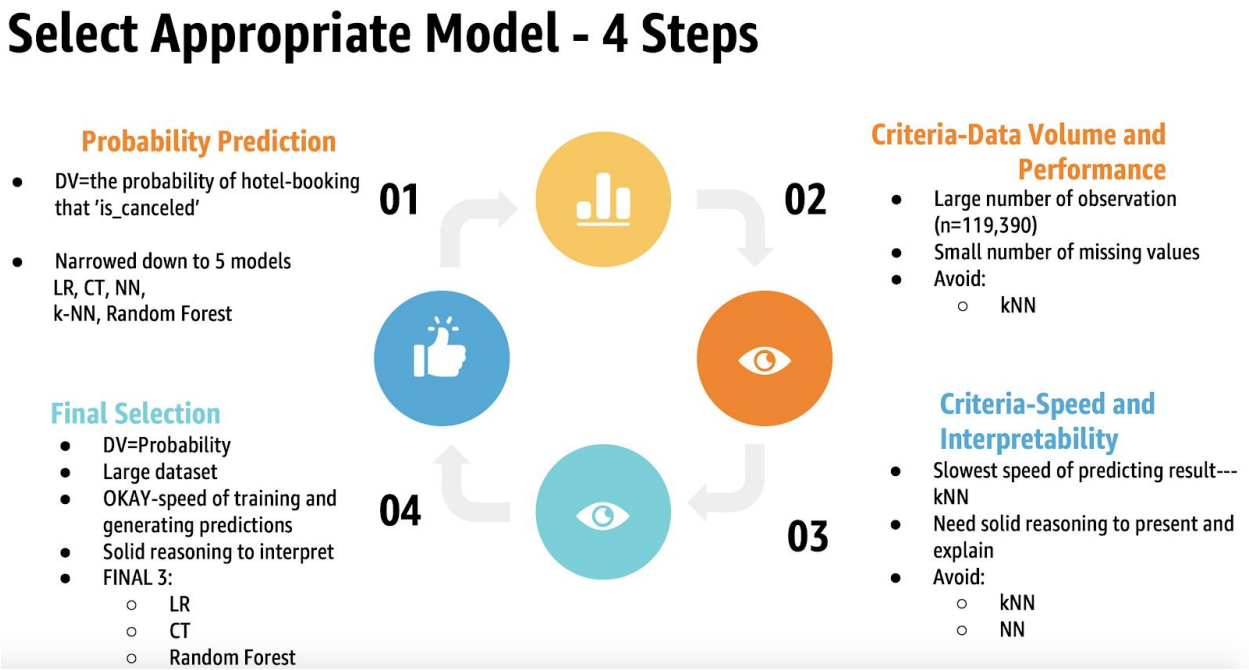
Regarding the categorical variable 'Special requests', more than 50% of hotel-bookings do not come with special requests. So we transformed this variable into 'made_special_requests' and 'not_made_special_requests'.



Ⅳ. Model Analysis

1. Model Selection

We selected models following the guidelines given in Session 12. We considered the criterias of data volume, model performance, speed of training model and generating predictions, and the interpretability to decide which technique we would adopt.



2. Results of Modeling

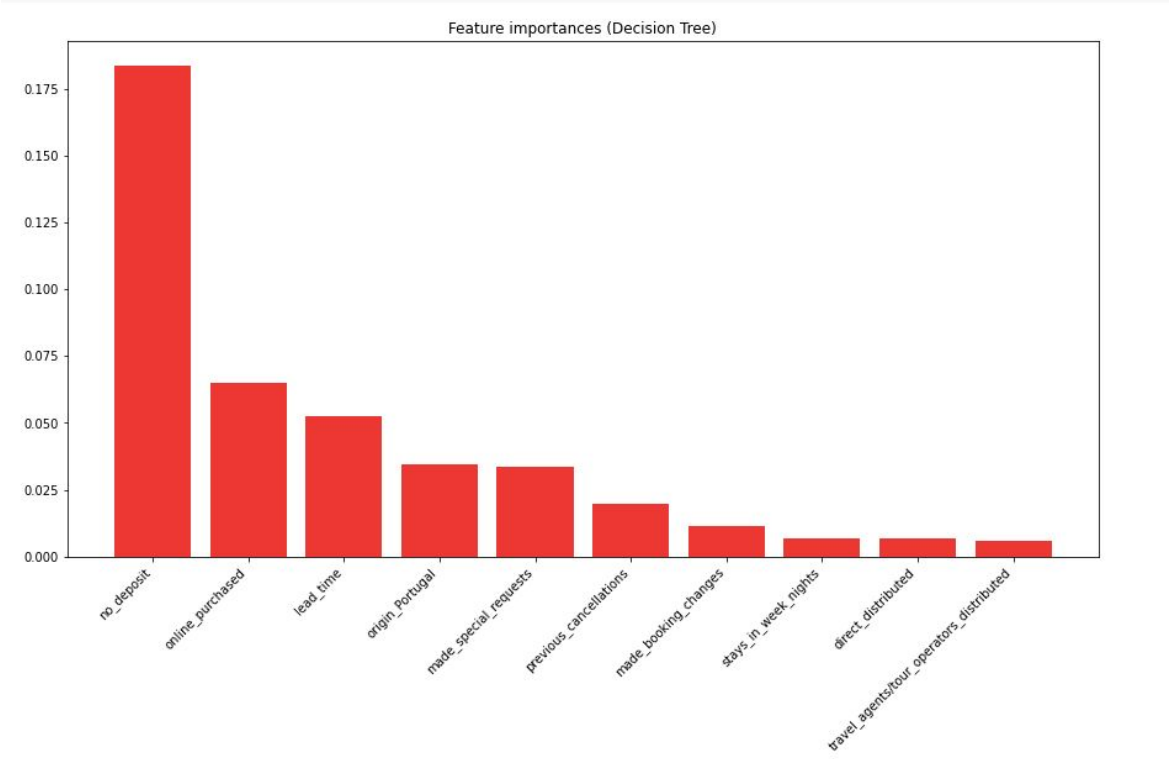
After using data from the test partition, we got really high ROC_AUC scores among all the models. Of these three, the Classification Tree was the most successful one with the largest Area Under ROC Curve, close to 90%. The Logistic Regression and Random Forest model have ROC_AUC rates of 87% and 89%, respectively.

Area Under Curve (AUC) of Each Models

Logistic Regression	0.8671642726600972
Random Forest	0.8901145714559978
Classification Tree	0.8959226024016617

Since this ROC_AUC rate was high, we decided to dive in and make some assumptions about why this was the case. The main thing that stood out to us was that the main area of focus for the Random Forest was whether or not the customer had put down a deposit. This made sense because the deposit was the largest contributor to all three methods, and it seemed to be the largest for the Random Forest.

As we have two numeric features and 17 categorical features, we decided to run a feature importance ranking on our most accurate model: decision tree. The top 3 features include no deposit, online purchase, and lead time. Since all of the coefficient of these factors are positive. We find three strong positive correlations: no hotel booking deposit leads to a better chance of cancellation; online purchased hotel booking leads to a better chance of cancellation; the longer time between booking and arrival, the better chance of cancellation.



The Classification Tree has produced the most accurate result, whose level of depth of the Classification Tree is 10, and the total number of Leaf Nodes is 390. The top 3 English rules and observation are shown as below:

TOP 3 English Rules (Classification Tree)

Reasoning: satisfied two criteria simultaneously

- Significance-large number of supporting observations
- Effectiveness-high probability

#TOP 1

Node ID 49 (most effective)

- The number of supporting observations is 9895 (close to one third of the cancellation)
- Probability is 100%

```
Leaf node ID = 49
Path = ['with_chilren <= 0.5', 'previous_cancellations > 0.5', 'no_deposit > 0.5', 'arrival_middle_season > 0.5', 'come_alone <= 0.5']
sample = 9895
value = [0, 9895]
class =  1
```

- English rule

IF ‘with_children=0’ AND ‘previous_cancellations’=1 AND ‘no_deposit’=1 AND ‘arrival_middle_season’=1 AND ‘come_alone’=0,

THEN PREDICT 1

#TOP 2

Node ID 224

- Number of supporting observations is 5132
- Probability is 99.6%

```
Leaf node ID = 224
Path = ['with_chilren > 0.5', 'is_canceled > 8.5', 'previous_cancellations <= 0.5', 'direct_distributed <= 0.5', 'travel_agents/tour_operators_distributed <= 0.5', 'is_repeated_guest <= 0.5', 'stays_in_weekend_nights > 1.5', 'no_deposit > 0.5', 'made_booking_changes <= 0.5', 'is_canceled <= 207.5']
sample = 5132
value = [5112, 20]
class =  0
```

- English rule

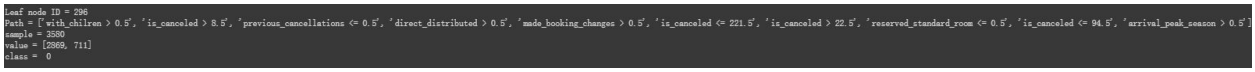
IF ‘with_children=1’ AND ‘previous_cancellations’=0 AND ‘direct_distributed’=0 AND ‘travel_agents/tour_operators_distributed’=0 AND ‘is_repeated_guest’=0 AND ‘stays_in_weekend_nights>=2’ AND ‘no_deposit’=1 AND ‘make_booking_changes’=0,

THEN PREDICT 0

#TOP 3

Node ID 296

- Number of supporting observations is 3580
- Probability is 80.13%



- English rule

IF ‘with_children’=1 AND ‘previous_cancelation’=0 AND ‘direct_distributed’=1 AND ‘made_booking_changes’=1 AND ‘reserved_standard_room’=0 AND ‘arrival_peak_season’=1,

THEN PREDICT 0

3. Observations and Management Insights

We wanted to highlight our top-performing node in particular because its number of supporting observations totals 9895, which makes up close to one-third of the cancelations. Because of this, it makes it the most critical key out of the three we are highlighting. It is even more important to note that the probability of this particular node is 100%. The accuracy and probability of this particular node are staggering, and we can assume that this node, in particular, influenced all three of our models.

We can also conclude the reason why the number of observations is so high is that the inclusion of ‘no_deposits’ and ‘previous_cancelation”. These were two of the largest contributors to each of the models.

Hotel Profit Calculation

Hotel Room / Customer	Not Prepared	Prepared
Cancel booking(s)	0	-2
Check-in	8	10

We used the above decision metrics to train prediction models to maximize the average net profit. For hotels, preparing a guest room will not always bring a profit: an unexpected cancellation will bring a loss of $2. On the other hand, a customer without cancellation always brings a profit. It is better to get the reserved room prepared with the information that a customer tends to show up and check-in. Facing the room reservation that is positively being canceled by a customer, the hotel should not arrange resources to prepare the room or find alternatives, lowering the cost by $2.

According to the decision metrics, we needed to calculate the decision cut-off to train the model. We made the decision rule based on the logic that we better not prepare the room and probably find the alternative if the customer is highly likely to cancel the booking(s). Since we code the cancellation as 1, our decision rules are based on the probability of cancellation:

IF Predicted Prob (Cancel) > X, THEN Not Prepare, ELSE Prepare

The profit calculation based on the probability of cancellation is as follows:

IF profit_Prepare > profit_Not_Prepare, THEN hotel will choose to prepare

ELSE hotel will choose not to prepare

After comparing the profit based on X, we calculate the decision cut-off as 0.5.[1]

Profit Calculation of Each Models

	Average Net Profit	Total Profit
Logistic Regression	$5.88	$702392.5
Random Forest	$5.92	$706630
Classification Tree	$5.98	$714047.5

After training the models with cross-validation for a better profit, we get our results shown above. The best-performing model is from Classification Tree, which will generate $5.98 average net profit from each booking and $714047.5 total income for three years, which brings $111655 more profit than the logistic model regression generates.

Thus, our final selected model is Classification Tree. We can predict individual hotel cancellations using an “auc_roc” trained model to help decide if it is profitable to arrange resources (including human resources and other financial cost) to prepare the room for the booking reservation. This is the way we applied the Classification Tree prediction model to improve decision-making power and eventually achieve the Advantage and the business goal of minimizing the cost and maximizing profits.

Ⅴ. Evaluation & Summary

In this project, we built a DIDA framework to analyze this dataset and generated insights regarding which predictors are significant in predicting whether a customer will cancel the hotel reservation. We demonstrated how our prediction model could help hotels avoid putting resources in preparing the room for predicted-canceled reservation, decrease the cost and eventually increase the profits. Our prediction model's AUC is 0.89, which indicates the predicting performance is confirmed.

[1] Look for detailed calculation, see Appendix 1

We broke the question down into a set of binary questions, cleaned the dataset, and did description analysis. We used both institutional knowledge and correlation analysis to perform feature selection and checked whether the dataset meets requirements.

In Exploratory Analysis, we performed data visualization, including bar-chart, pie-chart, and map. We chose the top three popular methods in model analysis: Logistic Regression, Random Forest, and a Classification Tree. After using data from the test partition, among all the models, the Classification Tree was the most successful/accurate with close to 90% ROC_AUC rate. We further analyzed each models' advantages and disadvantages.

Then we used decision metrics to train prediction models with cross-validation and calculate the decision cut-off to make the decision rule that maximizes the average net profit. Our final prediction model will generate $5.98 average net profit from each booking and $714047.5 total income for three years, and it also performs well with new data.

VI. Appendix

Decision cut-off calculation:

	Prepare the room	Not prepare the room
Actual 1 Cancel	-2	0
Actual 0 Show up	10	8

For the dependent variable, 1 – Cancel ; 0 – Show up

Thus, the decision rule will be:

IF Predicted Prob (Cancel) > X THEN Not Prepare ELSE Prepare

Expected profit if I choose not to prepare: profit_1 = p * 0 + (1-p) * 8 = 8 * (1-p)

Expected profit if I choose to prepare: profit_2 = p * (-2) + (1-p) * 10

I will choose to prepare if profit_2 > profit_1, otherwise I will choose not to prepare

Solve for p from the following

p * (-2) + (1-p) * 10 > 8 * (1-p) => p < 0.5

Thus, our decision cut-off is 0.5. IF Predicted Prob (Cancel) > 0.5 THEN Not Prepare ELSE Prepare