

Hotel Cancellation Problem and Overbooking Tactics Analysis

MGT4187

Business Analytics Report

Group: 钮祜禄葫芦娃队

Contents

Hotel Cancellation Problem and Overbooking Tactics Analysis.....	1
1. Introduction & Problem Identification	3
1.1 Hotel Background Information.....	3
1.2 Industry overview -- overbooking as an industry tactic	3
1.3 Problem identification -- the high cancelation rate and No-deposit policy	4
2. Data Inspection and Exploration	4
2.1 Dataset description	4
2.2 Data cleaning.....	4
2.3 EDA.....	5
3. Attempt One —— Optimization of Overbooking	6
3.1 Methodology	6
3.2 Model Selection and Result.....	7
3.3 Recommendation.....	9
4. Attempt Two —— Reduction of Cancellation Rate	9
4.1 Methodology	9
4.2 Results	12
4.3 Recommendation - Clustering and Logistic Regression	15
4. Discussion	17
5. Conclusion	17
Reference.....	18

1. Introduction & Problem Identification

1.1 Hotel Background Information

This project dataset contains data information retrieved from 2015 to 2017, a holiday hotel and a city hotel. Both hotels are located in Portugal. The first hotel is located in the resort area of the city of Algarve. The second hotel is located in downtown Lisbon, the capital. The two hotels are 280 kilometers away by car, and both are located on the North Atlantic coast.

With beyond-average hospitality, this full-service hotel company sets a price level at around 100 euros and provides services for customers from over 18 countries. Most of the orders come from Portugal itself and its European neighbors such as the United Kingdom, France, Spain, and Germany. These hotels thus provide reservations both in the city and resort. Each year July to November is the peak of hotel passenger flow.

Their annual booking orders are more than 40 thousand orders and they serve at least 100 thousand customers per year. More importantly, 42% of customers had to book 3 months ahead of time, which indicates its popularity. For the former hotel, it had 7 out of 12 months achieved full occupancy and the other one achieved 10 out of 12.

1.2 Industry overview -- overbooking as an industry tactic

Overbooking means that the hotel accepts more reservations than actual availability and anticipates some customers will cancel. This strategy targets no-shows and last-minute cancellations.

Consider research published in the Journal of Applied Sciences, which concluded that, on an average night, hotels experience a no-show rate of between five and 15 percent. To protect against a vacant room scenario, a hotel might overbook its rooms by up to 15% (Global, n.d.). It is quite often seen in the industry that to maximize profits, some hotels use this model to over accept orders. Industry insiders see it as a mixed blessing; overbooking can be a double-edged sword, on one hand, it can create a backup plan for a canceled reservation to achieve full occupancy because if implemented without the right care, it can lead to long-term economic and reputational losses.

Booking cancellations adversely affect hotels' ability to accurately forecast their occupancy and occupancy revenue levels. That is why hotels adopt overbooking practices and more stringent cancellation policies.

1.3 Problem identification -- the high cancelation rate and No-deposit policy

From 2015 to 2017, this hotel cancelation rate has been around 63% and is much higher than the industry average rate which is 24%. In the hotel industry, booking cancelation can recast hotel occupancy and reduce revenue levels.

This may also relate to the hotel running with the "No-Deposit" Booking policy. From the dataset analysis, among all the cancelation cases, 88% of customers are in the "No deposit policy" type. According to Lodging Magazine (Mandelbaum, 2019), the revenue that hotels received for cancellations and no-shows increased by almost 12 percent annually on average from 2012 to 2016. It is reasonable for us to consider a recharge deposit and refund policy to deal with the problem.

As requested by the hotel which wants to be able to build a model that predicts whether customers will actually stay, we need to conduct predicting cancellations model which will help hotels reduce the costs of no-shows and gain potential revenues. To guide our following discussion, we put forward three questions: (1) What is the most suitable overbooking rate to reduce vacancy? (2) Can overbooking perfectly solve the cancelation problem? (3) How can any other options tackle the high cancelation problem?

2. Data Inspection and Exploration

2.1 Dataset description

The dataset used in this report, named "Hotel Booking Demand", contains 32 columns, including our dependent binary variable "is_canceled" and other 31 attributes describing customers' demographic and behavioral features as well as details of their orders. It comprehends 119,390 observations of orders with the expected arrival date between the 1st of July 2015 and the 31st of August 2017 from both the city and the resort hotel. Detailed descriptions of each feature are attached in the appendix.

2.2 Data cleaning

The dataset was processed preliminary with several steps of data cleaning processes.

We first checked the missing values. The number of null values in the features "country", "company", and "agent" are respectively 370, 84469, and 12221. Therefore, the variable

"company" is deleted due to a large number of missing values. Orders with no country recorded are assumed to be from the mode of the rest of the data points ("PRT" which means Portugal). And a new class labeled 0 is created for the feature "agent" (index of the agent company) to fill in the missing value in this column as the missing implies no agent participated in the order.

The repeated data generated from collection error was then washed off the repeated data generated from collection error, after which the number of observations left is around 80,000.

We also performed the data transformation process. The String type of data is transformed into numerical data. The original numerical data is processed with $\log(x+1)$ transformation to weaken the impact of extreme values in further mining.

Moreover, the 4 "NA" value in the feature of "children" (indicating the number of children) is replaced with mode 0. Additionally, an extremely high value (€5400) and some negative values of the average daily price of the room ("adr") are deleted as the abnormal data is estimated to be aroused from data collection error.

2.3 EDA

Next, exploratory data analysis is conducted to obtain a primary understanding of the cancellation problem.

1) Figure 1 shows that city hotel faces more severe cancellation problem compared to resort hotel.

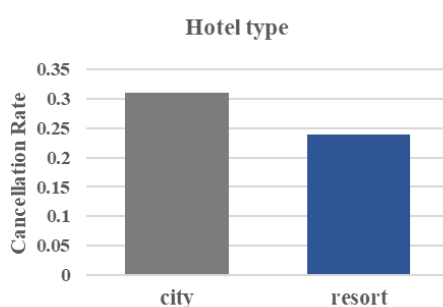


Figure 1. The cancellation rate of city hotel and resort hotel

2) The cancellation rate varies with the scheduled residence season as shown in figure 2. August sees the lowest cancellation rate of both hotels. The highest cancellation rate of the city hotel happens in January. And the resort hotel encounters the highest cancellation rate in June, the only month when the cancellation rate of the resort hotel is higher than that of the city hotel.

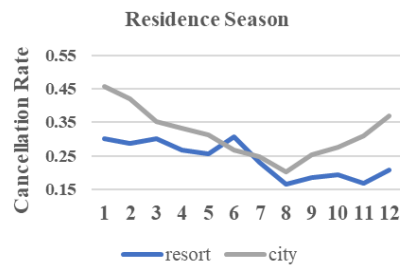


Figure 2. The cancelation rate and the residence season

3) Figure 3 indicates that first-time customers have a higher tendency to make cancellations compared to old customers.

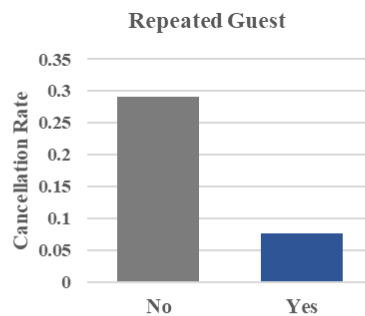


Figure 3. The cancellation rate of first time and old guests

4) A more special request is related to lower cancellations as shown in figure 4, which may be because satisfying the special needs of the guests can arouse their stickiness to the hotel.

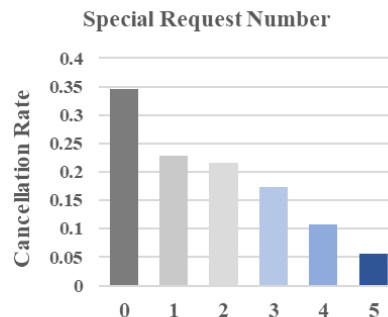


Figure 4. The cancelation rate and special request

3. Attempt One —— Optimization of Overbooking

3.1 Methodology

a. Feature Selection

Combine the result of EDA and various other feature selection methods, including ANOVA and linear regression, we select "whether_equal_room_type", "adults", "babies", "hotel", "meal", "is_repeated_guest", "lead_time", "previous_cancellations", "customer_type",

“required_car_parking_spaces”, “arrival_date_month”, “distribution_channel”, “agent”, and “deposit_type” as considered features.

b. Model

Firstly, we build and select a model which has the lowest total validation error rate. We try LDA, QDA, logistic regression, and random forest.

Secondly, to solve the potential imbalance between the false positive rate and false negative rate, we define security rate: $\{1 - [(actual\ shows\ up) / (actual\ no\ shows)]\}$ to evaluate the usage of room resources. If the security rate is smaller than zero, there will be some customers who cannot check-in, because the hotel does not reserve enough rooms for these customers. If the security rate is too large, there will be too many rooms reserved for the customer who has actually canceled the reservation. In that case, the hotel can maximize its room use ratio and not meet the situation that the customers cannot check in by making the security rate larger than zero and as small as possible. However, in the real world, the situation changes every day. So, we consider a 10% security rate as the best.

$$Security\ rate = 1 - (actual\ shows\ up) / (actual\ no\ shows) \quad (1)$$

We adjust the cutoff (i.e. predicted probability) to optimize the best model selected from the first step to have both a small total error rate and security rate close to 10%.

3.2 Model Selection and Result

We first build the random forest model and use it to show the importance of each feature.

```
rf.fit <- randomForest(is_canceled~whether_equal_room_type+adults+babies+hotel+meal+is_repeated_guest+lead_time+previous_cancellations+customer_type+required_car_parking_spaces+arrival_date_month+distribution_channel+agent+deposit_type,data=train_data[,importance=TRUE])
```

From the importance table, “deposit_type”, “previous_cancellations”, and “lead_time” are the top three most important features that mostly decrease the Gini index of the tree.

The importance of the feature is decreasing as the “MeanDecreaseAccuracy” decreases. We use the importance generated from random forest to fit the LDA, QDA, and logistic regression model by adding one feature each time from the most important feature to the less important feature.

	number_of_feature	error_lda	error_qda	error_log
[1,]	1	0.2646361	0.2646361	0.2646361
[2,]	2	0.2651254	0.2604009	0.2654771
[3,]	3	0.2687033	0.2602786	0.2682139
[4,]	4	0.2654618	0.2628319	0.2646973
[5,]	5	0.2618075	0.2614558	0.2612417
[6,]	6	0.2600491	0.2780297	0.2591317
[7,]	7	0.2589177	0.2779535	0.2591470
[8,]	8	0.2589789	0.3076312	0.2583980

Table 1. Cross-validation error of LDA, QDA, and logistic regression from 1 to 8 features

Comparing the cross-validation error rate, the random forest has the smallest error rate, which is 19.45%, so we further adjust the cutoff (i.e. predicted probability) on the random forest model to make the security rate close to 10%.

A cutoff (i.e. predicted probability) of 0.27 can make the security rate 13.52%, which is close to 10%. So, we chose random forest, cutoff (i.e. predicted probability) = 0.27 as our final model.

Cutoff	Actual shows up	Actual no shows	Security rate	Total error rate
0.05	34.8%	9.9%	-496.8%	25.53%
0.1	25.1%	15.1%	-181.0%	21.37%
0.2	17.0%	23.5%	-22.6%	19.41%
0.25	14.8%	26.4%	4.6%	19.11%
0.27	14.0%	27.5%	13.5%	19.03%
0.3	23.3%	28.7%	21.4%	19.03%
0.4	9.7%	35.1%	52.8%	19.13%
0.5	7.8%	39.3%	66.4%	19.45%

Table 2. Security rate and total error rate of models with different cutoffs

		Real	
		Positive	Negative
Predicted	Positive	TP 24,022	FP 7,892
	Negative	FN 9,126	TN 48,370

Table 3. The cross-validation confusion matrix of the final model

The test set was used to test the model. The model performs well with the test set accuracy rate equals to 78.87%. However, the false negative rate is 47.27%, which is unexpectedly high. The security rate is also much higher than expected, which means there will be some waste of room resources. This unexpected result may be due to the high variance of data itself.

Based on the test set, the cancellation rate is predicted to be 23.2%.

		Real	
		Positive	Negative
Predicted	Positive	TP 5,827	FP 1,089
	Negative	FN 5,224	TN 17,660

Table 4. The test set confusion matrix

3.3 Recommendation

In application, the overbooking system will not set a specific overbooking rate. It will use the registered customers' data to estimate each day's cancellation, then this part of the rooms can be overbooked. For example, if the hotel has 100 rooms, 60 of them are already booked for a specific day. The system will run the data of the 60 booking records and estimate that 10 rooms will be canceled. Then there will be 50 rooms available to be booked. 40 of them are not booked rooms and 10 are overbooking rooms. Based on the test set, the result of a 23.2% cancellation rate can be understood in this way – if the rooms are all booked, the hotel can have an overbooking rate of 23.2%. It means the model can help the hotel to raise its revenue from renting rooms or say, recover its loss due to no-shows by about 23.2% when it is running at full capacity.

4. Attempt Two —— Reduction of Cancellation Rate

4.1 Methodology

While the random forest model enables the hotel to calculate the overbooking rate to cover the revenue lost in the short term, the unstable result due to the high variance of the customer cancellation data requires us to tackle the problem from the root cause. The hotel needs to figure out the primary factors that contribute to the cancellation rate and design the marketing and management strategy based on these factors. In ideal circumstances, the result may both directly reduce the cancellation rate and indirectly improve the accuracy of the calculation of the random forest model. Logistic Regression and K-means clustering are used for modeling the data, which correspondingly find out the primary factors in the dataset and divide the customers into various sectors through unsupervised machine learning.

1) Logistic Regression

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. Logistic regression does not require a linear relationship between inputs and output

variables. This is due to applying a nonlinear log transformation to the odds ratio. Using the logistic function, the outcome of the model will be bounded in 0 and 1, which is well-fitted with the definition of probability.

$$\text{Logistic function} = \frac{1}{1 + e^{-x}} \quad (2)$$

Different from the random forest in that the result is hard to interpret, classification with the logistic regression can clearly show how the factors influence the cancellation rate through their coefficients. This enables the primary factors to be found. Furthermore, the model has various good properties. It is robust to small and medium noise of data and is not affected by slight multicollinearity. The possibility of classification is directly modeled without the need to assume the data distribution in advance, which avoids the problem caused by the inaccurate assumption of distribution. These properties are all suitable for this dataset that contains both numerical and categorical data.

As the long-term goal is to reduce the cancellation rate, the logistic model will be the best model to identify the primary factors. Before doing the logistic regression, it is necessary to first deal with the categorical data, as the categorical data which is assigned with category number in this dataset has no direct numerical ranking meaning in this case. Using the one-hot encoding, the model can use (n-1) dummy variables to represent n categories and the regression result can reveal the effect of each different category on the cancellation rate.

Order Number	Distribution Channel		Order Number	Direct	GDS	TA/TO	Undefined
001	1		001	0	0	0	0
002	2		002	1	0	0	0
003	4		003	0	0	1	0
004	1		004	0	0	0	0
005	3		005	0	1	0	0
006	5		006	0	0	0	1

Table 5. Example of one-hot encoding

All variables are considered in the logistic regression model and below is the formula for the model. The coefficients of the regression represent the effects of the factors and will be mainly focused on.

$$P(\text{Cancellation}) = \frac{e^{\beta_0 + \beta_1 \text{LeadTime} + \beta_2 \text{PreviousCancel} + \dots + \beta_k \text{Dummy}_{\text{DistributionChannel1}} + \dots}}{1 + e^{\beta_0 + \beta_1 \text{LeadTime} + \beta_2 \text{PreviousCancel} + \dots + \beta_k \text{Dummy}_{\text{DistributionChannel1}} + \dots}} \quad (3)$$

2) K-means clustering

As the logistic regression will identify the primary factors in the dataset, it is necessary to turn

data analysis into a business decision. It is less likely for the hotel to solve the problems for every primary factor, as it will be too time and money-consuming. To tackle this problem, the potential strategy proposed for our client is the stratified plan.

Clustering is the technique for finding subgroups, or clusters, in a data set when the observations within a group are similar but between groups are very different. K-means clustering starts with the first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. Finally, every point in the dataset will be assigned to k groups.

$$\text{minimize}\left\{\sum_{k=1}^K WCV(C_k)\right\} \quad (4)$$

$$WCV(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \quad (5)$$

Euclidean distance is used to represent the distance between the points and the centroids. The three most important factors will be chosen to stratify the customers and design corresponding strategies for each sector.

To choose the optimal number for the number of groups, the elbow method will be used. It works by finding the Within-Cluster Sum of Square which is the sum of the square distance between points in a cluster and the cluster centroid.

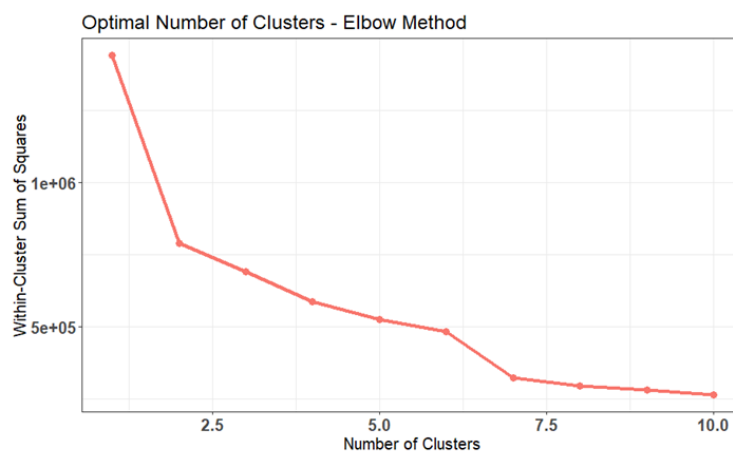


Figure 5. Elbow-Method

4.2 Results

a. Logistic Regression

Through the results of the logistic regression, 13 strongly statistically significant factors are picked out for further observation.

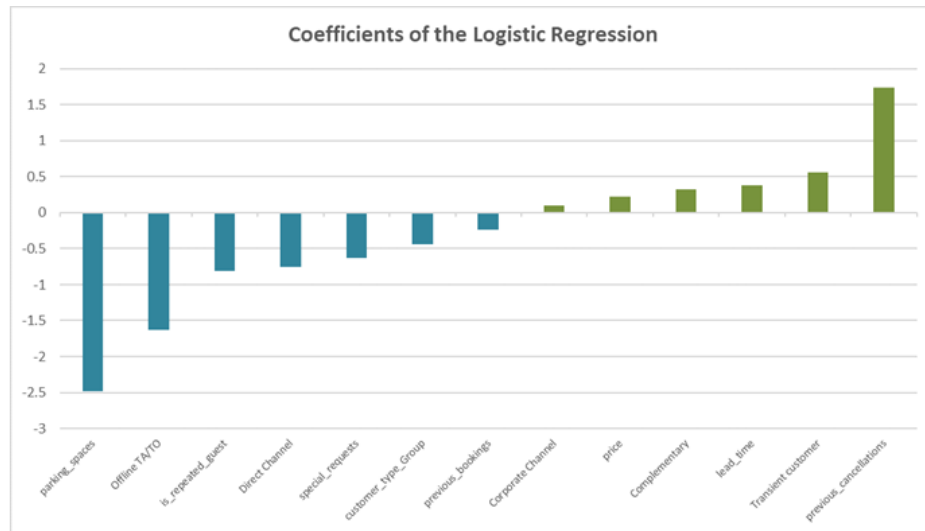


Figure 6. Results of Logistic Regression

Factors	Coefficients	Significance	P-value
parking_spaces	-2.478	***	<0.01
Offline TA/TO	-1.624	***	<0.01
is_repeated_guest	-0.810	***	<0.01
Direct Channel	-0.754	***	<0.01
special_requests	-0.626	***	<0.01
customer_type_Group	-0.435	***	<0.01
previous_bookings	-0.240	***	<0.01
Corporate Channel	0.100	***	<0.01
price	0.219	***	<0.01
Complementary	0.323	***	<0.01
lead_time	0.376	***	<0.01
Transient customer	0.555	***	<0.01
previous_cancellations	1.744	***	<0.01

Table 6. Coefficients of factors

The green factors in the above chart represent the factors that have a negative relationship with the cancellation rate, while the yellow ones may increase the cancellation rate. The absolute size of the numbers in the chart doesn't necessarily show the ranking of their influential power on the cancellation rate, as some categorical factors can only take a value of 1 or 0, however, numerical factors can take more values.

- 1) Positively related factors: Corporate Channel, Price, Lead Time, Previous Cancellations, Complementary Customer, Transient Customer.
- 2) Negatively related factors: Parking Spaces, Offline TA/TO, Repeated Guests, Direct

Channel, Special Requests, Group Customers, Previous Booking not Cancelled.

Though the model sacrifices some of the predicting accuracies to have higher interpretability, the predicting accuracy of the model is still satisfying, reaching an accuracy of 0.79. This shows that the model can find the correct relationship between the cancellation rate and the factors in this model.

```
Accuracy Score of Logistic Regression is : 0.7940376197799598
Confusion Matrix :
[[11184 1016]
 [ 2466 2240]]
Classification Report :
              precision    recall  f1-score   support

     0       0.82         0.92         0.87    12200
     1       0.69         0.48         0.56     4706

 accuracy         0.79         0.79         0.79    16906
 macro avg       0.75         0.70         0.71    16906
 weighted avg    0.78         0.79         0.78    16906
```

Table 7. Accuracy of Logistic Regression

b. K-means clustering

Constrained by the business purpose, it is necessary to further select the factors that are customer-behavior-related, which are more convenient for the hotel to design the strategies. Considering both the results of the random forest and logistic regression, three factors are selected for the following reasons.

Lead time is the most influential factor in both classification models, Previous bookings not canceled can show the dependency and frequency of the customer behavior and the price factor represents the monetary behavior. Using these three factors, every order will be assigned to specific groups and the hotel will design strategies for them.

Elbow Method is used for the clustering, and it is suggested that three groups will be the optimal number of groups for this model. Below is the result of the elbow methods.

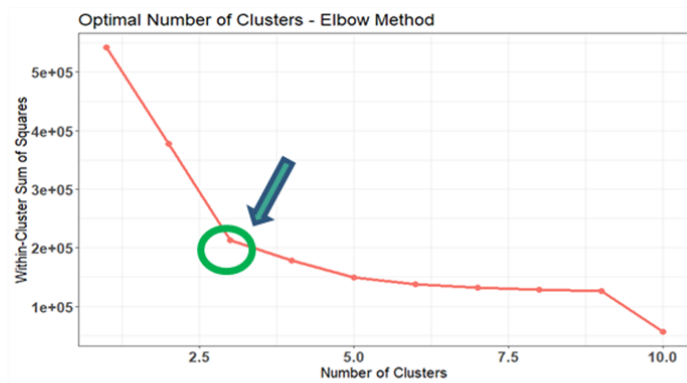


Figure 7. Result of Elbow Method

Then, the customer order data is divided into three groups. The plot and the average factor data of the customers are shown below.

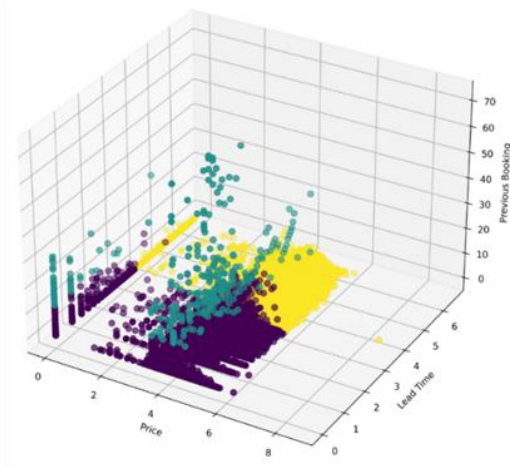


Figure 8. Result of clustering

Sector	Cancel rate	Number of customers	Lead time	Previous Bookings not Canceled	Price
1	0.13	26119	1.46	0.29	4.27
2	0.06	319	1.58	22.76	3.56
3	0.35	58089	4.45	0.02	4.61

Table 8. The average factor data of customers

Combining the table and the plot, it can be known that sector 1 is the purple part with a 13% cancellation rate, sector 2 is the green part with a 6% cancellation rate, while sector 3 belongs to the yellow part with a 35% cancellation rate.

Customers in sectors 1 and 2 have significantly lower cancellation rates than those in sector 3. It can be concluded as below.

- 1) High lead time has a strong positive relationship with the cancellation rate.
- 2) Customers with higher frequency result in a lower cancellation rate.
- 3) Price is also positively related to the cancellation rate. Higher price, higher rate.

Sector 1 and Sector 3 account for the main part of the customers, while Sector 2 has a lower population. All these statistics above will contribute to the business strategies in the recommendation part.

4.3 Recommendation - Clustering and Logistic Regression

a. Recommendation from Logistic Regression Result

As discussed previously, the random forest model can help accurately estimate the cancellation rate and set a red line for overbooking operations so that the downside of this tactic can be under control. However, although the model does help to remedy the revenue lost due to cancellation, the percentage that can be covered is unstable. Therefore, we would like to supplement the overbooking strategy with cancellation prevention measures.

Based on the insights gained from logistic regression, several initiatives can be proposed to prevent cancellation.

First, as the factor of lead time has a significant and positive coefficient, the hotel may want to pay special attention to customers booked long before their stay. The customized direct push should be delivered to them via email and message after their booking periodically. Contents of the pushes can include reminder messages of their scheduled residence and the services that are ready for them to enjoy, ads of the hotel's featured service, festival and birthday greetings, etc.

Secondly, since cancellation possibility varies with distribution channel and customer type, the hotel may want to make corresponding adjustments in terms of distribution strategy and marketing target. It can reserve more rooms for its direct channel, which has a negative coefficient of -0.8 in the logistic regression but is utilized in only 12% of its orders currently. Since being a group customer is also negatively correlated with cancellation possibility, the hotel may make focused marketing on the group and contract customers, including designing group discounts and selecting marketing channels that have higher exposure to group customers to deliver their ads.

Thirdly, as the number of special requests being satisfied has a negative coefficient in the logistic regression, the hotel can conduct marketing research on customers' demand for special needs and enrich the types of special services the hotel offers.

The above initiatives can be developed into mature strategies based on the hotel's exact condition and preference.

b. Recommendation from Clustering Result

By comparing our client with the industry data (TripSavvy, 2020), it is common for hotels,

especially luxury hotels, to make an advanced deposit. One potential problem the hotel may have is the "deposit plan". From the data, 88% of reservations are of no deposit, so it is very likely to result in a high cancellation due to "No extra expense" for customers.

To tackle this problem, the potential strategy proposed for the hotel is the "Stratified deposit plan". According to industry experience (Aims, R., 2022), the deposit is one of the effective approaches to lower the cancellation rate. Thus, the hotel can assign different deposits for different clusters of customers to achieve the goal of cancellation reduction.

As a result of clustering, according to cancellation levels from low to high, customers can be divided into three different clusters (Diamond, Golden, Copper). For a Diamond customer, their deposit can be waived because of trust and great historical performance; For a Golden customer, they need to pay 50% room fee as a deposit; For the Copper customer, 70% of the room fee as a deposit. By doing this, the cancellation of part of the customers can be avoided, especially those Copper customers who are more likely to cancel.

One thing to note is that the deposit is refundable as long as the cancellation is 48 hours (2 days) ahead of the check-in time. There are several reasons behind this:

- (1) Customer satisfaction won't be influenced much by the new "stratified deposit plan".
- (2) The main goal of the "stratified deposit plan" is to avoid last-minute cancellations. In another word, if a customer cancels a long time ahead of check-in day, the hotel can rearrange the room by overbooking optimization model and its popularity among customers (Reserve 3 months ahead) as mentioned in the background information.

Besides 3 key variables, we also consider other related external factors, which are Month & Hotel Type. According to the graph, there is an obvious seasonal difference in cancellation rate (i.e. low in summer & fall, high in winter & spring). For hotel type, the cancellation rate is low for resort type & high for city type. Therefore, it is suggested another 5% deduction in the deposit respectively for customers who book from July to November and those who book a Resort hotel. It is a kind of motivation for high-credibility customers to book again next time without cancellation.

Together with clustering, a stratified deposit plan for the hotel is comprehensive and our clients can lower the cancellation rate by implementing this strategy.

4. Discussion

The above two attempts are expected to enable the hotel to largely mitigate its loss holistically caused by the cancellation. However, there are still certain concerns and limitations, and some need to be addressed by further study.

First, in the random forest model, as we adjust the cutoff with the purpose to mitigate the risk of overly aggressive overbooking, optimizing the confusion matrix performance is not our priority. Also, considering the high variation of the data used and the nature of the cancellation problem, the accuracy of the predicted cancellation rate, especially the false negative rate, can be unstable.

In addition, with the implementation of supplementary measures that reduce the long-term cancellation rate, customers' tendency to cancel will change. We recommend that hotels adjust the forecasting models based on the real-time data they collected to regulate overbooking.

Moreover, the deposit rate in the stratified deposit plan needs to be made more solid and data-driven with continuous observation of more customer behavior data and measurement of the elasticity after implementing the current plan.

5. Conclusion

The problem of the high hotel cancellation rate was serious and caused financial losses to the hotel. We primarily recommend an overbooking policy to improve the utilization of customers' canceled rooms. As overbooking cannot perfectly solve all the losses and some vacancies are remaining, we recommend other strategies that tackle the root of customers' credit on cancellation to reduce the cancellation rate in the long term. One is a stratified deposit plan that sets different deposit discounts for customers. Another is more initiatives like an upgraded promotion mechanism adjusting the distribution channel, and more based on data pattern.

To work out a suitable overbooking rate, we selected 14 outstanding features and chose the random forest model to predict customers' cancellation rates. Based on our adjusted model which adopted a security rate of 10% to guarantee customers' successful check-in with reservations, we calculated an overbooking rate of 23.2%. Then, we conducted logistic regression and found the 10 most significant factors affecting the cancellation probability. We chose the lead time, price, and the previous booking not canceled as key attributes and set K equals 3, the elbow, to do clustering. We designed a stratified deposit plan based on the three

clustered customer groups.

Reference

Aims, R. (2022, April 9). How Do Hotel Security Deposits Work|Hotel Advice. The Alcazar.
<https://thealcazar.com/how-do-hotel-security-deposits-work>

Mandelbaum, R. (2019b, May 6). Showing No Shows: A Closer Look at Attrition and Cancellation Fee Revenue. LODGING Magazine. <https://lodgingmagazine.com/showing-no-shows-a-closer-look-at-attrition-and-cancellation-fee-revenue/>

Global, S. (n.d.). *Overbooking Your Hotel? How To Do It The Right Way*.

<https://hsb.shr.global/learning-center/overbooking-howtodoittherightway>

Mandelbaum, R. (2019, May 6). *Showing No Shows: A Closer Look at Attrition and Cancellation Fee Revenue*. LODGING Magazine.

<https://lodgingmagazine.com/showing-no-shows-a-closer-look-at-attrition-and-cancellation-fee-revenue/>

Reserving a Room: Advance Deposits. (2020, August 6). TripSavvy.

<https://www.tripsavvy.com/advance-deposit-1895484>