Research on the Influencing Factors of Cancellation of Hotel Reservations

Yaqi Lin *

Department of Automation and electrical engineering, Jinan University, Shandong, 271000, China

* Corresponding author: 202030308005@stu.ujn.edu.cn

Abstract. Booking hotels online is now a very common way for people to travel and stay, but a large number of cancellations due to itinerary changes and other factors can have a big impact on hotels, such as losing customers who really need a certain room type and losing them to other hotels. In order to reduce hotel losses, this paper uses the data of two hotels through data published on Kaggle's official website, identifies the factors that have the greatest impact on hotel cancellations through EDA visualization, and gives improvement measures. Machine learning algorithms are then used to guess whether the customer will cancel the booking. Each algorithm has its own area of expertise, so this article makes a comparison to the performance of decision trees, logistic regression, random forests. The result is that random forests have the highest accuracy and hotel managers can use the model to predict and change business strategies to increase profits.

Keywords: Booking cancellation, EDA, machine learning.

1. Introduction

Booking a hotel in the current society is a common way to make sure you can stay in the room you want. In the Internet era, the concept of revenue management was first proposed by the aviation industry [1]. It has been used in many fields such as hotel management. In the hotel industry, in order to allocate more resources to customers, many hotels have adopted online booking to increase hotel awareness and make it easier for customers to book rooms. However, due to external uncertainties such as flight cancellations, itinerary changes, conference cancellations, many customers choose to cancel their orders after booking a room.

Therefore, this article needs to understand what causes customers to cancel and reduce hotel losses by building models to predict which part of customers will cancel. The customer churn theory [2] is when a customer breaks off working with the current company and instead works with another more advantageous company. In the hotel field, it is that the customer cancels the reservation due to the objective conditions of the hotel rather than uncertain external factors and books a room in another hotel [3]. It is pointed out that there are many factors that affect the choice and change of reservations, but they are not responsible for it, and there will be some impact on the hotel [4]. So this paper analyzes the factors of customer churn through data and predicts customer churn based on the model and find the most appropriate model through the degree of fit, which is the purpose of our research.

Therefore, based on this demand, people use different machine model requirements in order to predict hotel cancellations. In the beginning, most of the forecasting models were used in the aviation industry [5]. Antoniod showed that knowledge of data visualization, machine learning, and statistics can successfully predict order cancellations [6]. Since then, many people have also become interested in the application of machine learning algorithms to the field of hotel reservations. Neslin and Gupta found that logistic regression models and decision trees were good predictors of customer churn [7]. Caigny propose a method for predicting hotel cancellations through a hybrid model of logistic regression and decision trees [8]. Antonio, et al. used a machine learning classification model to predict hotel cancellations [9]. Jasmina Novakovic1 and Snezana Turina compared different machine algorithms to predict order cancellations [10]. Antonio analyzed the data of four hotels and successfully predicted cancellations, allowing hotels to maximize efficiency based on the number of bookings given by the model [10].

In this study, this paper will analyze the data that has been made public, find out why customers cancel orders and how they change, and predict the cancellation of reservations through different modeling methods.

2. Method

2.1. Data Description

Data used in this article is an open dataset downloaded from Kaggle website. Since it is real information taken, in which the real information of all customers is eliminated, people can use it with peace of mind. All information was provided by Jesse Mostipak three years ago. The following table provides a detailed description of these data variables, as Table 1 shows.

Table 1. Variable interpretation

Name	Description
hotel	Resort Hotel or City Hotel
Is canceled	Value indicating if the booking was canceled or not
lead_time	Number of days between reservation and check-in date
arrival_year	Time of year of arrival
arrival_month	Date of monthly arrival
arrival_week_number	Weekly number of the year for the date of arrival
arrival_day_of_month	Date of the day of entry
stays on weekends	Number of weekends the customer has stayed or booked to stay
stays on weeks	Number of overnight stays per week for which the customer has stayed or booked
adults	Amount of adults.
children	Amount of children
babies	Amount of babies
meals	Meals ID requested by guests
country	Which country the customer is from
market_segment	Market segmentation to which the booking was assigned
distribution_channel	The name of the distribution channel used for the booking
is_repeated_guest	Customers who have stayed at this hotel once or more
previous_cancellations	The number of times the customer canceled the order
previous_bookings_not_canceled	Number of reservations that the guest hasn't previously canceled
reserved_room_type	The room type that the customer needs
assigned_room_type	The room type assigned to the reservation
booking_changes	Intention to cancel the booking due to some factor
deposit_type	Whether a deposit is required before check-in
agent	ID of agent
company	ID of company
days_in_waiting_list	Days the reservation was on a waiting list before
customer_type	Type of client
adr	Daily average rate
required_car_parking_spaces	The quantity of parking places the visitor needs
total_of_special_requests	The quantity of unique requests made by customers
reservation_status	Current reservation status
reservation_status_date	Date of booking status

2.2. Exploratory Data Analysis (EDA)

Data cleansing, data description (description of statistics, charts), data distribution, data comparison, data intuition creation, data summary, etc. are among the main duties of EDA. It mainly includes the following three aspects: distribution analysis, statistical analysis and correlation analysis.

Distribution analysis means quantitative qualitative analysis. Statistical analysis is concentration, discrete trends and distribution shapes. Correlation analysis contains single plot, graph matrix, correlation coefficient. The data will be shown visually in this article along with the link between the variable and the goal variable, a list of variables that have a substantial impact on the target variable, and a model of the outcome.

2.3. Machine Algorithm

2.3.1. Decision tree

A decision tree approach can be used to approximation the value of a discrete function. It is a standard classification technique where the data is processed first, understandable rules, and decision trees are produced using inductive algorithms, and the new data is then assessed using the decisions. In essence, a decision tree is a method for classifying data based on a set of criteria.

2.3.2. Random forest

The approach is known as a random forest algorithm because it combines many decision trees, randomly inserts the dataset each time, and randomly chooses some attributes to be input. It is clear that the random forest algorithm uses a decision tree as an estimator and is a Bagging algorithm.

Bagging algorithm is an ensemble learning algorithm, its full name is self-aggregation algorithm (Bootstrap aggregating), as the name suggests the algorithm consists of Bootstrap and Aggregating two parts.

2.3.3. Logistic regression models

Through the Logistic function (Sigmoid function), which maps the data features to a probability value in the 0–1 interval (the probability that the sample belongs to the positive case), logistic regression, which is a type of linear classifier, determines the classification to which the data belongs by comparing with 0.5. In order to estimate the parameters for a logistic regression, one must first assume that the data conform to this distribution.

3. Results and Discussion

First, this article visualizes the data so that more directly can be determined which factors are the cause of hotel cancellations through graphs and tables. Due to the large amount of data, this paper randomly selected 50,000 pieces of data from 119391 data for analysis.

Figure 1 shows that 37% of all cancellations are attributable to the number of cancellations, and Figure 2 shows intuitively how many cancellations per hotel there were. The profile of cancellations at both hotels is high, but city hotels have a cancellation rate that exceeds that of resorts by about 12%. So, what are the key factors that influence customer cancellations? Information can be get by comparing each piece of data.

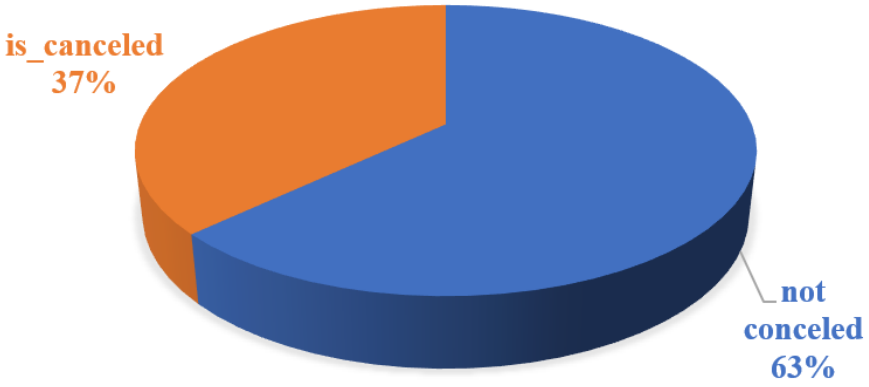

Figure 1. The rate of cancellations and no-cancellations

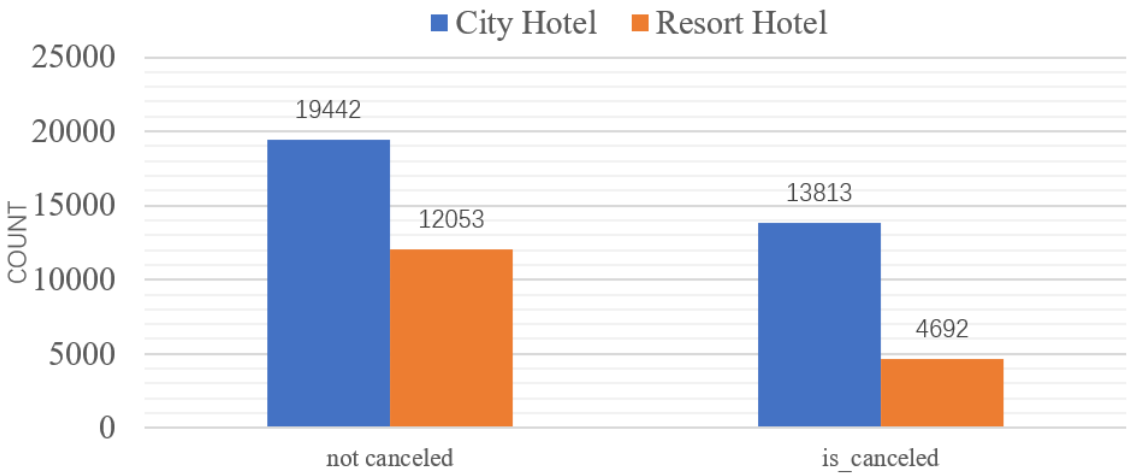

Figure 2. Number of cancellations and non-cancellations by hotels

In Figure 3, this article will analyze whether the number of cancellations due to other factors changes due to different months in terms of time. Intuitively see that July and August are the peak period for customer orders, and it is also the maximum number of canceled orders. Therefore, the increase in the number of tourists in July and August also led to many customers canceling orders for various reasons. But it can also find that even if the number of reservations is much smaller in winter, the cancellation rate of reservations will also be less than in summer, so the number of cancellations in summer hotels in winter will be reduced. That's why the managers can offer big discounts in the summer and open up more rooms for bookings.

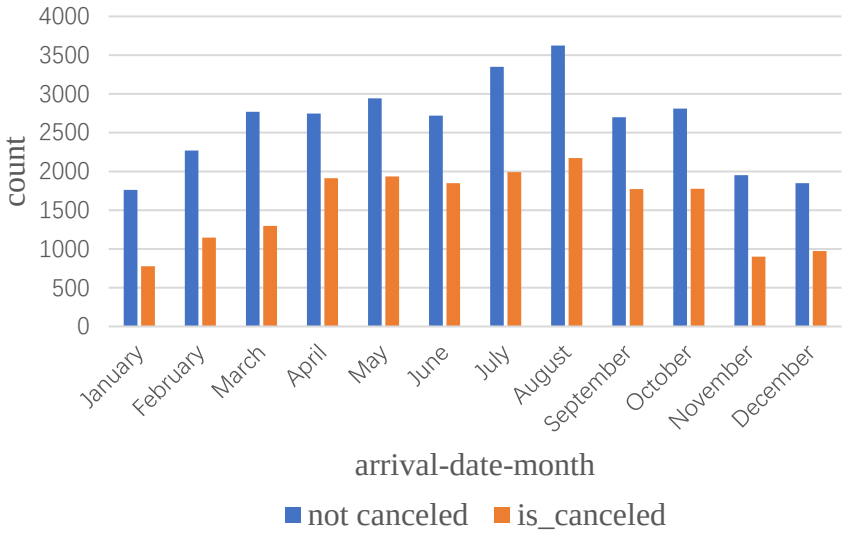

Figure 3. The number of orders cancelled or not cancelled in different months

According to the analysis of the length of advance booking time and cancellation rate, in Figure 4 it can clearly see that the length of advance booking is positively correlated with the cancellation rate, and the likelihood of a cancellation increases with the length of advance reservations. Yet, the likelihood of cancellation decreases as check-in time approaches. Therefore, for some popular room types, managers can have limited reservations within one month to reduce cancellations.

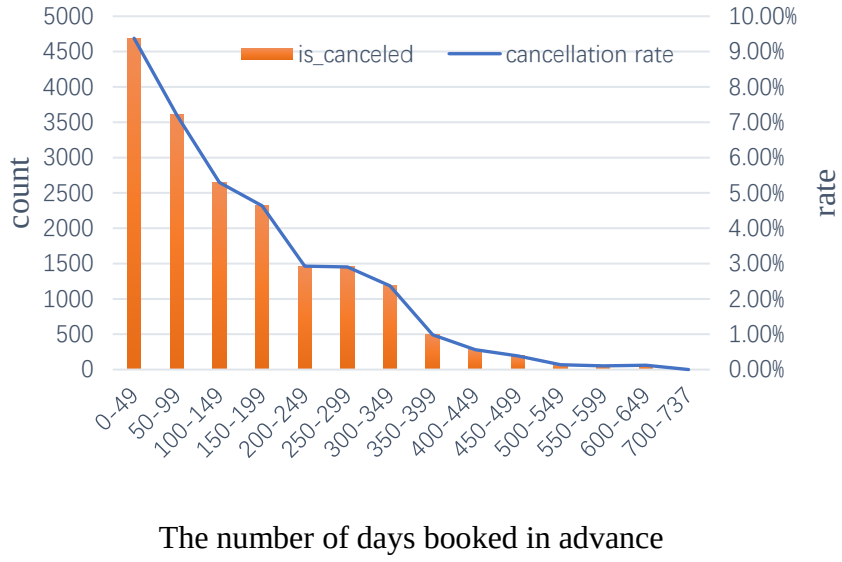

Figure 4. The length of advance booking time and cancellation rate

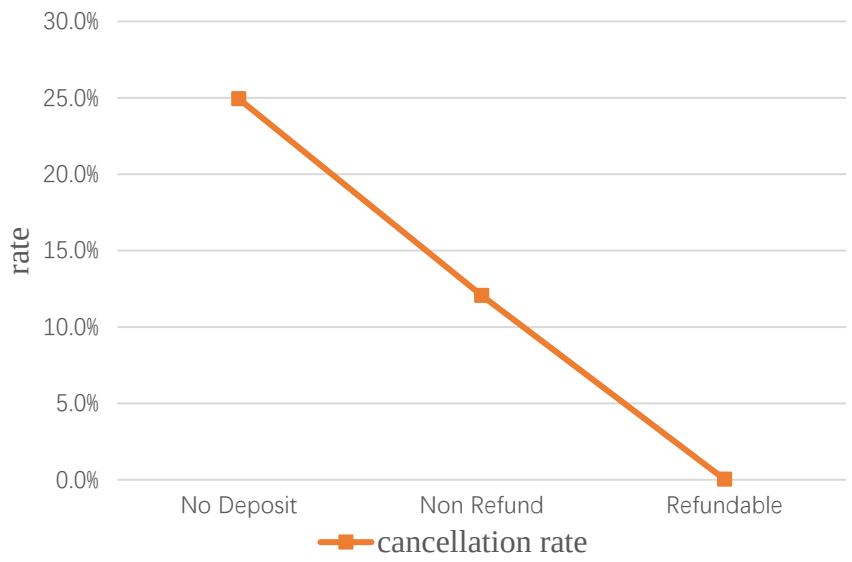

Figure 5. Cancellation rate

Then look at whether the payment of the deposit will affect the cancellation of the hotel reservation. Both hotels basically do not pay a deposit, but it is clear in Figure 5 that the cancellation rate of no deposit is the highest, so take the deposit method to reduce the practice of customers booking rooms at will, but it may also increase customer churn. Therefore, deposit can be changed according to the room type. Then a statistic needs to be made about the type of room. From Figure 6, it is clear that Type A rooms are the most booked, but have the highest cancellation rates. So, managers can reduce cancellations by charging a deposit for popular room types.

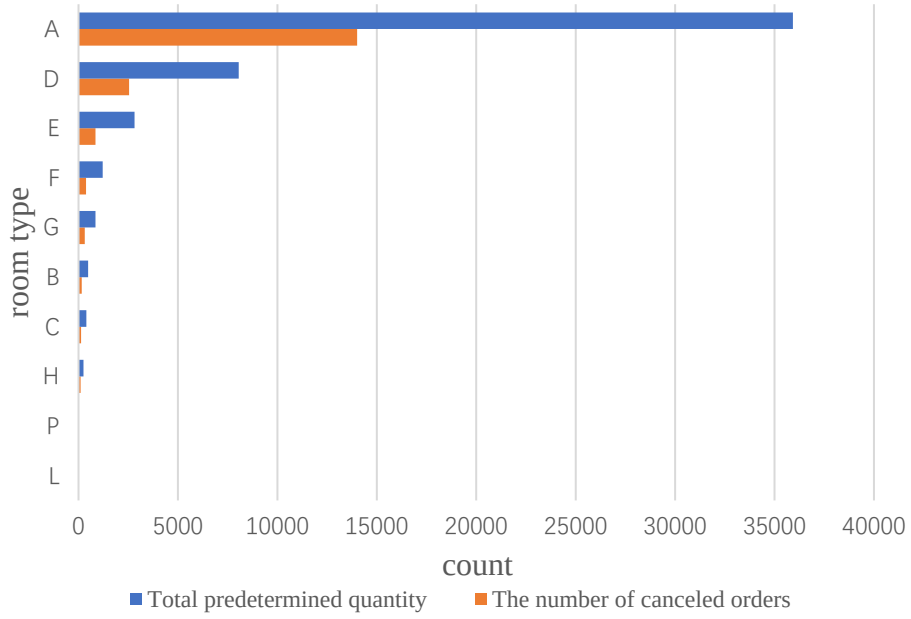

Figure 6. Number of different rooms and cancellation rate

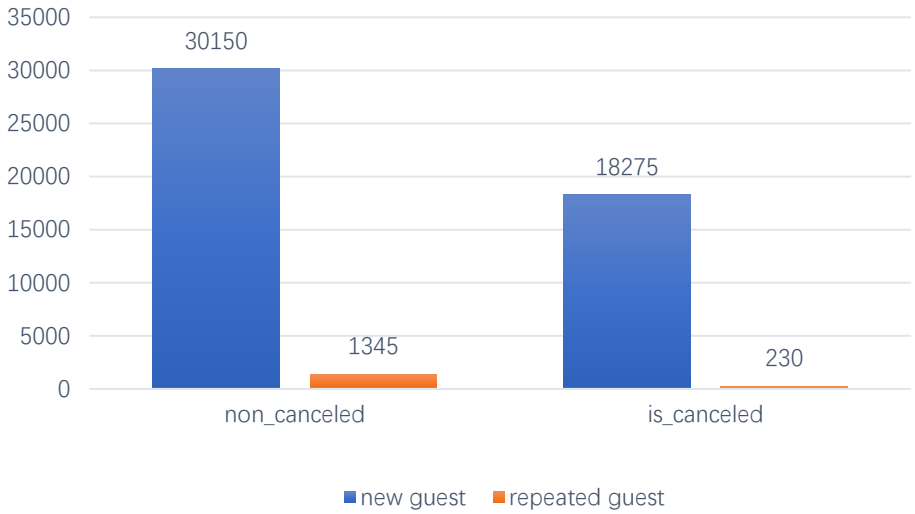

Figure 7. Reservations made by repeated guests and new guests, cancelled and not cancelled

Next, analyze whether the customer will go on to book the hotel the person has stayed at before, and how the cancellation rate differs from the new customer. This can be seen in Figure 7 that the chances of repeated customers to book are very small, but the cancellation rate is low. So, hotels can attract old customers to book this hotel again by taking discounts, no deposits, etc. That can also let them play a good publicity role.

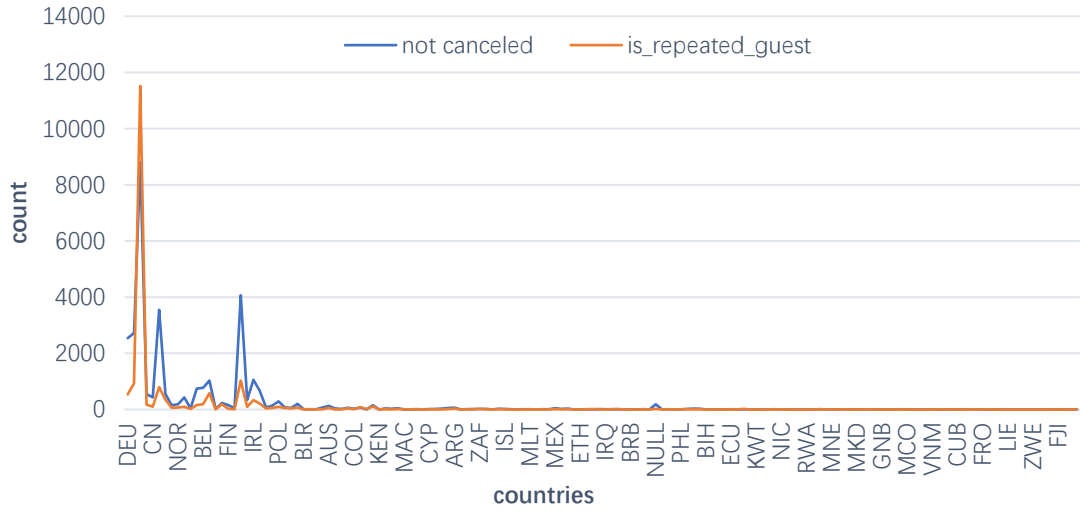

Figure 8. The number of cancellations by customers from different countries

The article counted the nationalities of clients in Figure 8 and discovered that the cancellations were primarily from Europe, proving the influence of geography on hotel cancellation rates. Since these two hotels are located in Portugal, Europe also has the largest number of tourists and it is also the reason why it has large cancellation rates.

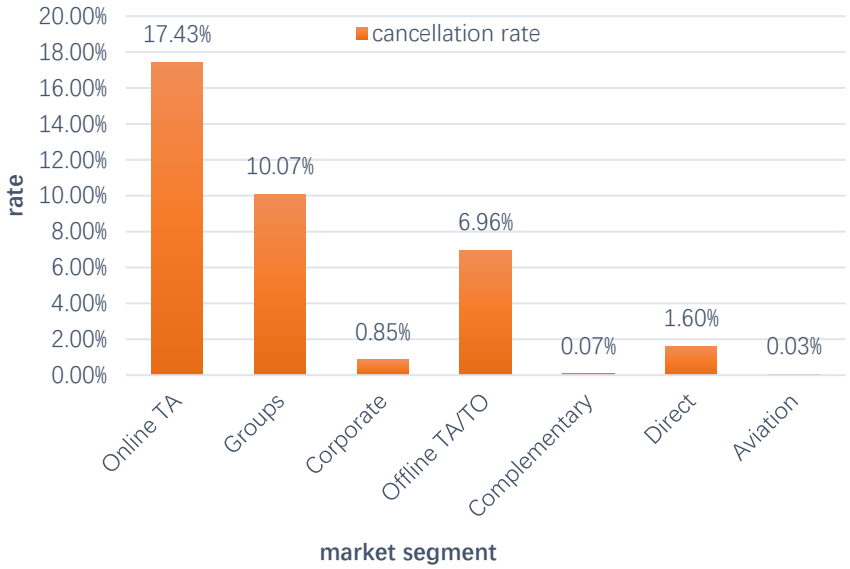

Figure 9. Market segment and cancellation rate

Then this paper analyses which groups have the highest number of unsubscribes from the perspective of booking groups. Figure 9 shows that the largest rate of unsubscribes is from Online TA. Article can pay more attention to this travel agency, or build a list based on the number of cancellations to prevent the loss caused by excessive cancellations.

Figure 10. The number of rooms with and without cancellations by the city hotel in different adr

Figure 11. The number of rooms with and without cancellations by the resort hotel in different adr

This article compares which type of room has the most cancellations based on hotel price. Comparing the lines in Figure 10 and Figure 11, it is clear that the likelihood that a reservation would be canceled increases with price. Relatively speaking, the less likely customers are to cancel their booking if the price is less than 100 euros.

Table 2. Parameter Estimates (*n*=50000)

	Unstandardized Coefficients		Standardized Coefficients	*t*	*p*	VIF
	B	S.E.	*Beta*			
Constant	-0.833	0.016	-	-53.58	0.000**	-
lead_time	0.001	0	0.142	34.464	0.000**	1.197
hotel	-0.06	0.004	-0.059	-14.448	0.000**	1.162
arrival_month	-0.004	0.001	-0.031	-8.145	0.000**	1.023
country	0.002	0	0.196	47.458	0.000**	1.213
market_segment	0.058	0.002	0.153	36.857	0.000**	1.214
is_repeated_guest	-0.062	0.011	-0.022	-5.641	0.000**	1.103
reserved_room_type	0.007	0.001	0.024	5.551	0.000**	1.374
deposit_type	0.554	0.006	0.382	85.86	0.000**	1.399
adr	0.001	0	0.06	13.723	0.000**	1.361
R	0.294					
Adj *R*	0.294					
F	*F* (9,49990)=2311.543,*p*=0.000					
The value of D-W	1.993					
Dependent Variable: is_canceled						
* *p*<0.05 ** *p*<0.01						

Finally, according to linear regression, the size of the correlation coefficient is used to determine which factors can most affect the customer's cancellation of the hotel. Lead time, hotel, arrival month, country, market segment, is repeated guest, reserved room type, deposit type, adr were used as independent variables in the linear regression analysis, and is canceled was the dependent variable, as shown in table 2.

All of the VIF values are less than 5, which the model's test for multicollinearity indicates means there is no collinearity problem. Also, the D-W value is near to 2, indicating that the model is more accurate because the sample data and the variables in the model do not auto-correlate or have any association.

If the p-value in the preceding figure is significant (0.05 or 0.01), it means that X has an effect on Y. The complete examination of the data thus reveals that the lead time's regression coefficient value is 0.001 (t=34.464, p=0.0000.01), indicating that the lead time will significantly improve the is canceled. By analogy, lead_time, country, market_segment, reserved_room_type, deposit_type, adr will have a significant positive impact on is_canceled. Furthermore, the variables hotel, arrival month, and is repeated guest can significantly degrade the value of is canceled. So, in order to reduce the impact of these factors on hotel cancellations, steps should take to reduce these things.

Most of the customers are new customers, and only 3.15% of old customers, but the cancellation rate is very low. Hotels need to consider upgrading in terms of returning customers. For example, you can increase the repurchase rate of customers by launching a membership system, issuing coupons and other activities.

Winter is when city hotels and resorts receive the fewest reservations and charge some of the lowest prices of the year. However, year-round city hotels and resorts are the most popular for A room types and have higher occupancy rates. Hotels can develop more personalized marketing campaigns for these two types in different seasons to increase user booking and occupancy.

In terms of booking methods, most customers use online booking to book hotels. Hotels can try more marketing activities in online ticket booking to increase the conversion rate of customers. In addition, some travel agencies with high cancellation rates have been adopted to establish cancellation lists to reduce the loss caused by a large number of room reservation cancellations

In terms of cancellation rate, city hotels do not fluctuate much from month to month, while resort hotels are more affected by the season, with a high cancellation rate of large orders in summer and a relatively small cancellation rate in winter. Characteristics closely related to reservation cancellation include the type of user's prepaid deposit, the length of advance booking, etc., and hotels need to focus on these characteristics and take corresponding countermeasures, such as proactive intervention by calling customers in advance to minimize the occurrence of temporary cancellation by customers. At the same time, hotels should make full use of existing data to build a prediction model of whether customers will cancel orders, predict whether customers will cancel orders, formulate hotel room allocation plans within the predicted number of canceled orders, adjust operation strategies in a timely manner, and minimize the impact of customer cancellations on hotel revenue and other aspects.

Table 3. Categorization decision tree model

Decision Tree	Precision	Recall	f1-score	sample
0	0.82	0.83	0.83	6319
1	0.7	0.7	0.7	3681
accuracy			0.78	10000
macro avg	0.76	0.76	0.76	10000
weighted avg	0.78	0.78	0.78	10000

Table 3 demonstrates that because reservations constitute the majority of what hotels offer, reservations information must be made in advance. If the hotel's reservation is canceled, there can be unrecognized issues, thus management need to anticipate the cancellations in order to identify any issues before they arise. Data modeling is then necessary. To create standard data, feature extraction is done first. Due to the different unit dimensions of numerical features, it is easy for the model to be biased during fitting, therefore normalization processing, unification of the dimensions, and retention of the data rules are required. The sample characteristics and sample results are then divided. Finally, the model is built and evaluated. In this paper, decision trees, logistic regression and immediate forest models with a wide range of known applications are selected, and the models are fitted respectively, and the results are shown in Tables 3, 4, and 5 below.

The higher the accuracy, which is measured as the percentage of samples with accurate prediction findings in relation to the whole sample, the better. Precision, which is essentially the proportion of positive samples, is the proportion of positive outcomes anticipated by the outcome. The higher the recall, or the percentage of positive samples that were expected to be positive, the better. The

harmonic average of accuracy and memory, or F1-score, is a complete evaluation indicator that encompasses both precision and recall.

The better the accuracy and recall rate, but as these two factors are sometimes at odds with one another, the f1-score is frequently employed to assess the classifier's performance in its whole. Its value range is 0 to 1, with 1 being ideal. But the specific should be combined with the actual situation to evaluate. It is often used to evaluate the improvement of a model.

Table 4. Categorization model using random forest

random forest	precision	Recall	f1-score	sample
0	0.81	0.87	0.84	6319
1	0.74	0.65	0.69	3681
accuracy			0.78	10000
macro avg	0.77	0.76	0.76	10000
weighted avg	0.78	0.78	0.78	10000

Table 5. Logistic Regression

		Predicted		Predicted Correct	Predicted Error
		0	1		
Observed	0.0(n=31495)	29722	1773	94.37%	5.63%
	1.0(n=18505)	10449	8056	43.53%	56.47%
Overall				75.56%	24.44%

So it can clearly find that logistic regression is the weakest, and random forests show the highest accuracy results.

4. Conclusion

This article predicts consumer cancellation of hotel reservations using EDA, machine learning models, etc., and identifies the elements that influence customer cancellation. Visual data processing and scientific model prediction make it possible to predict behavior, and there are different prediction results for different hotels and different influencing factors.

Therefore, in order to reduce the loss caused by a large number of hotel reservation cancellations, hotel managers need to use the model to run every day to predict today's cancellation probability and then determine the number of room types open today, etc., and arrange rooms reasonably to maximize profits.

References

[1] Kimes S E, Wirtz J. Has revenue management become acceptable? Findings from an Interna-tional study on the perceived fairness of rate fences. Journal of Service Research, 2003, 6 (2): 125 – 135.

[2] Hadden J, Tiwari A, Roy R, et al. Computer assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research, 2007, 34 (10): 2902 - 2917.

[3] Bai Ruirui. Research on customer prediction and retention countermeasures of hotel reservation platform. Zhengzhou University, 2021.

[4] Talluri K T, Van R G. The theory and practice of revenue management. New York, NY: Springer, 2005.

[5] Ahlam A, et al. Cancellation Prediction for Flight Data Using Machine Learning. 2nd International Conference on Advances in Science & Technology (ICAST), 2019.

[6] Antonio N, et al. Using data science to predict hotel booking cancellations. Handbook of Research on Holistic Optimization Techniques in the Hospitality Tourism and Travel Industry. Hershey, PA, USA: Business Science Reference, 2016, 141 - 167.

[7] Neslin S, Gupta S, Kamakura W, et al. Defection detection: improving predictive accuracy of customer churn models. Tuck School of Business, Dartmouth College, 2004.

[8] De Caigny A, Coussement K, De Bock K W. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 2018, 269 (2): 760 - 772.

[9] Antonio N, Almeida A, Nunes L. Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model. 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, 2017, 1049 - 1054.

[10] Jasmina Novakovic1, Snezana Turina. Hotel reservation cancellations: analysis and prediction using machine learning algorithms. International academic journal, 2021.