

Peer review

Lau, Grant, and Abdulla Kerimov. 2025. "Hotel Reservation Cancellations: Predictive Modeling and Feature Impact Analysis." *Journal of High School Science* 9 (1): 303–20.

1. This model does not seem to be of much use in achieving the result of preventing the non-utilization of rooms due to cancellations. A probability of cancellation of X rooms for a certain date or time (or for a certain period of time) does not enable the hotel operator to gross overbook (because the probabilities calculated are not specific to the separate categories of # of adults, the type of reservation (room type), # of week or weekend nights, the market segment type or the month of booking.....) Hence, the information obtained - regardless of how close to 1 the F1, precision or recall values are - is not actionable. The author should have segregated the probability of cancellation or non-cancellation to a specific categorical variable or variables, the data so obtained would then be actionable. Therefore, all claims to F1 scores, precision or recall being better than earlier literature are meaningless. I do not anticipate that this manuscript will be acceptable without addressing this challenge of actionability.
2. In the context of point 1, reference #4 by Chen is hence actionable because the author integrates the public dataset with a PNR (person, name, records) dataset so that the model predicts the probability with which a certain individual will cancel a certain reservation.
3. Provide evidence through correlograms or a collinearity matrix (with Pearson's coefficients) that the first 6 influential variables do not exhibit collinearity. This can further be quantified by VIF.
4. Many graphs do not have units. For example, average price of room, lead time.. Please check for all missing units and include these.
5. The training/test data split is not presented in the manuscript. Was K-fold validation performed on the data? If not, why not? Over how many epochs was the model trained (when relevant)? What was the loss to begin with and at the end of the epochs? Why was no validation performed?
6. SHAP analysis assumes feature independence. You have not provided evidence that your first six features (in order of importance to the model's accuracy) are independent of each other (see point 3). In this context, using SHAP for feature importance becomes circular reasoning and is error prone.

This model does not seem to be of much use in achieving the result of preventing the non-utilization of rooms due to cancellations. A probability of cancellation of X rooms for a certain date or time (or for a certain period of time) does not enable the hotel operator to gross overbook (because the probabilities calculated are not specific to the separate categories of # of adults, the type of reservation (room type), # of week or weekend nights, the market segment type or the month of booking.....) Hence, the information obtained - regardless of how close to 1 the F1, precision or recall values are - is not actionable. The author should have segregated the probability of cancellation or non-cancellation to a specific categorical variable or variables, the data so obtained would then be actionable. Therefore, all claims to F1 scores, precision or recall being better than earlier literature are meaningless. I do not anticipate that this manuscript will be acceptable without addressing this challenge of actionability.

Thank you for your detailed feedback. We would like to clarify that our model is designed to predict the probability of cancellation for a specific reservation based on the given input features, as outlined in Table 1 (we recently added Table 1). The input variables—such as reservation type, arrival month, market segment type, number of adults and children, number of week and weekend nights, and other features—are used by the model to generate a probability of cancellation for that particular booking. We acknowledge the reviewer's concern regarding actionability and the potential need for segmenting probabilities based on categorical variables. However, our model inherently provides actionable insights by allowing hotel operators to assess the likelihood of cancellation for any individual booking.

This information can support targeted strategies, such as adjusting overbooking policies dynamically based on predicted cancellation risk, rather than relying solely on aggregated probabilities.

Furthermore, our evaluation metrics, including F1-score, precision, and recall, remain meaningful in this context. These metrics reflect the model's ability to correctly classify cancellations and non-cancellations at the individual reservation level. While prior studies may have reported different methodologies, our approach focuses on predicting cancellations with high accuracy for specific reservations rather than aggregated trends.

If needed, we can further clarify this aspect in the manuscript to ensure that the practical implications of our model are clearly conveyed. Please let us know if additional modifications would be beneficial.

1. In the context of point 1, reference #4 by Chen is hence actionable because the author integrates the public dataset with a PNR (person, name, records) dataset so that the model predicts the probability with which a certain individual will cancel a certain reservation.

Please see question 1.

2. Provide evidence through correlograms or a collinearity matrix (with Pearson's coefficients) that the first 6 influential variables do not exhibit collinearity. This can further be quantified by VIF.

We sincerely thank the reviewer for their constructive feedback. In response to the reviewer's suggestion, we have included a Pearson correlation matrix, which can be found in Figure 2 along with its description in the 'Exploratory Data Analysis' section. Upon review, we did not observe any highly correlated features nor any indication of multicollinearity

We used Pearson correlation to calculate the correlation coefficients between the variables. As seen in Figure 2, 0.54 and 0.47 are the largest correlation coefficients between the number of previous bookings not canceled and repeated guests pair, and the number of previous bookings canceled and not canceled pairs, respectively. Because there are no highly correlated variables with coefficients larger than ± 0.8 , we conclude that there is no evidence of colinearity present in our dataset.

3. Many graphs do not have units. For example, average price of room, lead time. Please check for all missing units and include these.

Thank you very much for the comment. We have included units in Figures 7 and 8.

4. The training/test data split is not presented in the manuscript. Was K-fold validation performed on the data? If not, why not? Over how many epochs was the model trained (when relevant)? What was the loss to begin with and at the end of the epochs? Why was no validation performed?

We sincerely thank the reviewer for pointing this out. In response, we have added a detailed description of the train-test split in the 'Methods and Models' section. Additionally, we clarified that we used 3-fold cross-validation during hyperparameter tuning. As for the training process, epochs do not apply to our models.

Before model development, we separated the data into an 80-20 train-test split. Additionally, because our dataset is mildly imbalanced, we set stratify equal to "true." This keeps the ratio of non-canceled and canceled reservations the same across both the training and testing data. When hyperparameter tuning all our models using the train data and grid search method, we set cross-validation to 3 and the evaluation metric to weighted f1-score.

5. SHAP analysis assumes feature independence. You have not provided evidence that your first six features (in order of importance to the model's accuracy) are independent of each other (see point 3). In this context, using SHAP for feature importance becomes circular reasoning and is error prone.

We appreciate the reviewer's insightful comment. As noted in response to question 3, we have ensured that the features in question are not highly correlated and that there is no indication of collinearity.

Given this, we believe the assumption of feature independence holds for our model. Therefore, we do

not consider the use of SHAP for feature importance to involve circular reasoning or introduce error. We hope this clarification addresses the concern.

I am not sure that my questions 1 and 2 have been addressed. The Chen reference includes the PNR hence the hotel can cancel a certain individual's booking based on the ML model. My original questions still is unanswered "...A probability of cancellation of X rooms for a certain date or time (or for a certain period of time) does not enable the hotel operator to gross overbook (because the probabilities calculated are not specific to the separate categories of # of adults, the type of reservation (room type), # of week or weekend nights, the market segment type or the month of booking.....) Hence, the information obtained - regardless of how close to 1 the F1, precision or recall values are - is not actionable." Part of your response reads "...However, our model inherently provides actionable insights by allowing hotel operators to assess the likelihood of cancellation for any individual booking....." I don't understand how each individual booking can be assigned a probability of cancellation. Please explain in detail - perhaps using an example - in the manuscript. For example, if I were a repeat customer (2 adults and one child), booking 2 weekday nights in the month of June, with a preference for a double queen bed room on the top floor, not near the elevator, overlooking the courtyard or the pool, paying with points accumulated during my previous stays, with no previous cancellations, flying into town with 4 days lead booking time and needing 1 car parking space, can your algorithm predict the cancellation probability for this reservation? If not, since my hotel has 200+ rooms, how do i know which rooms will have a high cancellation probability? The manuscript's usability is therefore minimal regardless of the metrics.

I am not sure that my questions 1 and 2 have been addressed. The Chen reference includes the PNR hence the hotel can cancel a certain individual's booking based on the ML model. My original questions still is unanswered "...A probability of cancellation of X rooms for a certain date or time (or for a certain period of time) does not enable the hotel operator to gross overbook (because the probabilities calculated are not specific to the separate categories of # of adults, the type of reservation (room type), # of week or weekend nights, the market segment type or the month of booking.....) Hence, the information obtained - regardless of how close to 1 the F1, precision or recall values are - is not actionable." Part of your response reads "...However, our model inherently provides actionable insights by allowing hotel operators to assess the likelihood of cancellation for any individual booking....." I don't understand how each individual booking can be assigned a probability of cancellation. Please explain in detail - perhaps using an example - in the manuscript. For example, if I were a repeat customer (2 adults and one child), booking 2 weekday nights in the month of June, with a preference for a double queen bed room on the top floor, not near the elevator, overlooking the courtyard or the pool, paying with points accumulated during my previous stays, with no previous cancellations, flying into town with 4 days lead booking time and needing 1 car parking space, can your algorithm predict the cancellation probability for this reservation? If not, since my hotel has 200+ rooms, how do I know which rooms will have a high cancellation probability? The manuscript's usability is therefore minimal regardless of the metrics.

Thank you for giving us the opportunity to clarify our model's capabilities. We appreciate your insights and included a detailed example in the first paragraph of the 'Results and Discussion' section.

We would like to clarify it further. Our model does not predict the probability of cancellation for a specific date or time. It predicts whether a given booking will be canceled based on its individual characteristics. The model considers a set of booking input features, including:

- Number of adults and children
- Number of weekday and weekend nights
- Type of meal plan
- Required car parking space

- Room type reserved
- Lead time
- Arrival month
- Arrival day of the week
- Market segment type
- Whether the guest is a repeat customer
- Number of previous cancellations and non-cancellations
- Average price per room
- Number of special requests

Regarding the Chen reference, we would like to clarify that their ‘Computational study one’ also utilizes almost the same input features(data) as ours; see their Appendix A.

Table A1
Detailed information of the PNR data set in computational study one.

No.	Variable	Variable type	Statistical description
1	<i>is_canceled</i> (dependent variable, DV)	Categorical	2 categories: canceled and not canceled
2	<i>hotel</i>	Categorical	2 categories: resort hotel and city hotel
3	<i>lead_time</i>	Numerical	Number of lead days
4	<i>arrival_date_year</i>	Variable related to time	Arrival year: 2015–2017
5	<i>arrival_date_month</i>	Variable related to time	Arrival month: 1–12
6	<i>arrival_date_week_number</i>	Variable related to time	Arrival week: 1–53
7	<i>arrival_date_day_of_month</i>	Variable related to time	Arrival day in month: 1–31
8	<i>stays_in_weekend_nights</i>	Numerical	Number of weekend nights
9	<i>stays_in_week_nights</i>	Numerical	Number of week nights
10	<i>adults</i>	Numerical	Number of adults
11	<i>children</i>	Numerical	Number of children
12	<i>babies</i>	Numerical	Number of babies
13	<i>meal</i>	Categorical	4 categories: type of booked meal
14	<i>country</i>	Categorical	Customer source country
15	<i>market_segment</i>	Categorical	8 categories: market segment designation
16	<i>distribution_channel</i>	Categorical	5 categories: booking distribution channel
17	<i>is_repeated_guest</i>	Categorical	2 categories: whether is a repeat consumer
18	<i>previous_cancellations</i>	Numerical	Number of previous cancellations
19	<i>previous_bookings_not_canceled</i>	Numerical	Number of previous bookings not canceled
20	<i>reserved_room_type</i>	Categorical	9 categories: reserved room type
21	<i>booking_changes</i>	Numerical	Number of booking changes
22	<i>deposit_type</i>	Categorical	3 categories: deposit type
23	<i>agent</i>	Categorical	Travel agency ID
24	<i>days_in_waiting_list</i>	Numerical	Number of days in the waiting list
25	<i>customer_type</i>	Categorical	4 categories: type of booking
26	<i>adr</i>	Numerical	Number of average daily rate
27	<i>required_car_parking_spaces</i>	Numerical	Number of parking spaces
28	<i>total_of_special_requests</i>	Numerical	Number of special requests

Since our approach is a classification problem, the classification models output a probability of cancellation for each individual booking.

To answer reviewer’s question “For example, if I were a repeat customer (2 adults and one child), booking 2 weekday nights in the month of June, with a preference for a double queen bed room on the top floor, not near the elevator, overlooking the courtyard or the pool, paying with points accumulated during my previous stays, with no previous cancellations, flying into town with 4 days lead booking time and needing 1 car parking space, can your algorithm predict the cancellation probability for this reservation?” Yes, our model predicts this booking cancellation or probability of cancellation.

Thank you for your submission. The paper cannot be considered for publication until the following concerns are addressed:

Two existing papers, “Predicting Hotel Booking Cancellations using Machine Learning Techniques” (<https://ieeexplore.ieee.org/document/10725148>) and “Predicting Hotel Cancellations using Machine Learning: A Data Science Journey” (<https://medium.com/@ukpowehonome/predicting-hotel-cancellations-using-machine-learning-a-data-science-journey-43c27766ef07>), report F1 scores of 0.99 and 0.93, respectively, which are higher than those presented in your paper. Please provide a discussion comparing your results with these existing studies.

The second paper listed above identifies the top four influencing features as lead time, average price per room, number of special requests, and arrival date, while your paper lists lead time, number of special requests, market segment type (online), and average price per room. Given this difference, can your approach be generalized to all hotels, or is it applicable only to a specific subset? Please clarify. In Figure 7, the Y-axis should be labeled as “Lead Time (days),” and the sentence “Interestingly, more reservations with higher lead times” should be revised to “Interestingly, more reservations with higher median lead times.”

The statement “A threshold value is established by the model and used to decide which category to place a value.” requires further explanation. How is this threshold determined? Please elaborate.

The paragraph beginning with “(2 adults and one child), booking 2 weekday nights in the month of June...” should be removed, as it does not align with the formal tone of the paper.

In Table 3, the model “Tuned XGBoost + SMOTE Oversampling” appears to yield slightly better results than “Tuned Random Forest + Class_Weight: ‘Balanced’”, which the paper has chosen. Please provide a rationale for selecting this model over the alternative.

Can the model be further optimized to improve the F1 score for the “Cancelled” category? If so, please discuss potential avenues for improvement.

Reference #5 links to a page that is not found. Kindly update or replace this reference.

To enhance transparency and reproducibility, please consider making all relevant datasets available on GitHub or another publicly accessible repository.

Two existing papers, “Predicting Hotel Booking Cancellations using Machine Learning Techniques” (<https://ieeexplore.ieee.org/document/10725148>) and “Predicting Hotel Cancellations using Machine Learning: A Data Science Journey”

(<https://medium.com/@ukpowehonome/predicting-hotel-cancellations-using-machine-learning-a-data-science-journey-43c27766ef07>), report F1 scores of 0.99 and 0.93, respectively, which are higher than those presented in your paper. Please provide a discussion comparing your results with these existing studies.

While the reported F1 scores in the referenced papers appear higher, we would like to point out the following issues with those papers. In the first paper, the authors present an average F1 score despite the dataset being imbalanced, which can lead to inflated and potentially misleading conclusions. In contrast, our study reports F1 scores separately for each class to provide a clearer picture of model performance across different classes. Similarly, the second paper also reports an average F1 score and does not use a stratified split when dividing the dataset, which is inappropriate for imbalanced data. Stratification is essential to maintain class distribution in both training and test sets, and its omission may have led to misleading performance metrics.

Moreover, because both studies only report an overall F1 score, it is impossible to determine the F1 score for canceled and not-canceled reservations separately. It is likely that their models achieve a high score for the majority class (not-canceled) while performing poorly on the minority class (canceled), inflating the weighted F1 average without ensuring balanced predictive performance. In our study, we prioritize methodological rigor to ensure a more reliable evaluation of model performance.

We believe it is important to note that neither of these papers is peer-reviewed. We have not included this information in our introduction section. Does the reviewer kindly recommend including this background information in our introduction section?

6. The second paper listed above identifies the top four influencing features as lead time, average price per room, number of special requests, and arrival date, while your paper lists lead time, number of special requests, market segment type (online), and average price per room. Given this difference, can your approach be generalized to all hotels, or is it applicable only to a specific subset? Please clarify.
-

Thank you for addressing my comments. Accepted. Please review the attached galley proof for errors (see below) and return to us in 48 hours to ensure a timely publication.

I found one error and one omission.

1. You state "...If the probability exceeds the threshold, the model predicts that the input's category will be 1. If the probability is lower than the threshold, the model predicts the input's category will be 0. " Should the words "input's" be replaced with the word "outputs" ?
2. You said somewhere that you had made a Github account or equivalent as a repository for your models. Please provide a link so that we can include it in the manuscript.