Hotel Booking Cancellation Prediction

*IOE 473 Final Report

1st Ivana Zhao
Industrial and Operations Engineering
University of Michigan Ann Arbor
Ann Arbor, US
zhaozyf@umich.edu

2nd Zhonghui Wang
Industrial and Operations Engineering
University of Michigan Ann Arbor
Ann Arbor, US
zhonghuw@umich.edu

3rd Zeqing Zhou
Industrial and Operations Engineering
University of Michigan Ann Arbor
Ann Arbor, US
zeqingzh@umich.edu

***Abstract*—Room cancellation rate is an important aspect for hotel management since it could affect the final room occupancy rate and the hotel's revenue. In fact, very little is known about the reasons that customers cancel the reservation and how to lower the cancellation rate. The goal of this paper is to build a machine learning model to figure out what kind of customers are highly likely to cancel their bookings and predict the cancellation rate given customers' information. By training multiple classification models on historical data, we could choose the best model with the optimal performance to predict the cancellation rate. The prediction results can help hotel managers accurately forecast net demands, define better overbooking tactics and make other business strategies to keep customers and lower cancellation rates, and finally increase hotel profits.**

***Index Terms*—hotel cancellation, machine learning, prediction**

I. Introduction

A. Literature Review

In the hospitality industry, booking cancellations have a significant impact on the revenue and demand management decisions. Nuno Antonio et al. (2018) built a machine learning classification model to predict which bookings are "likely to cancel" and found that bookings contacted by hotels cancel less than bookings not contacted. Martin Falk et al. (2018) revealed that the role of booking lead time and country of residence of the customers in determining the cancellation probability is more pronounced for online than for offline or travel agency bookings.

B. Hypotheses

In this project, we employ a binary parameter called "is_canceled" to indicate whether a reservation has been canceled. We then examine the relationships between the cancellation likelihood and twenty independent variables, which represent the features of the bookings. The independent variables are presented in Table II.

Specifically, the rate of cancellations can differ based on the type of hotel, with city hotels being viewed as more of a temporary accommodation option while resort hotels are considered a vacation destination; longer lead time may result in a lower cancellation rate, as customers may have ample time to plan their travel arrangements and less likely to altering their plans; customers may face unforeseen circumstances that could result in cancellations during peak travel seasons, holidays, or when traveling with children; the type of meal included in the booking may also influence the cancellation rate because customers who have booked a non-refundable package that includes meals may be less likely to cancel their bookings; the traits associated with market segments and customer types can also impact the cancellation rate based on their respective characteristics; customers who have a history of canceling their bookings may be more likely to cancel again; clients who reserve a particular type of room may exhibit lower cancellation rates, provided they are not given an alternative room type; customers who make frequent changes to their booking may be more likely to cancel their bookings; the deposit type may also be a factor because customers who have made a non-refundable deposit may be less likely to cancel their bookings; an extended duration on the waiting list could raise the probability of cancellations, and the number of requests made may also impact the chances of a booking being canceled.

TABLE I
List of Independent Variables

Variable	Description
hotel_type	Hotel type
lead_time	Lead time
arrival_month	Arrival month
stays_in_week_nights	Stays in week nights
stays_in_weekend_nights	Stays in weekend nights
adults	Number of adults
children	Number of children
babies	Number of babies
meal	Meal
market_segment	Market segment
is_repeated_guest	Is repeated guest
previous_cancellations	Previous cancellations
reserved_room_type	Reserved room type
is_room_type_changed	Is room type changed
booking_changes	Booking changes
deposit_type	Deposit type
days_in_waiting_list	Days in waiting list
customer_type	Customer type
required_car_parking_spaces	Required car parking spaces
total_special_requests	Total special requests

C. Scope

The scope of the project is to build a classification machine learning model that can predict whether a future reservation would be canceled based on customer information. However, this project will only focus on analyzing the existing dataset and will not involve collecting new data. Furthermore, the aim of this project is not to examine particular hotels or locations, but instead to study a wide selection of hotels to detect commonalities and trends in the data.

D. Importance

By examining how customer characteristics affect hotel reservation cancellations, it is possible to make precise predictions that could enhance revenue management, customer contentment, resource distribution, and the industry's competitiveness, which would be advantageous for both hotels and their clients.

II. Data

We sourced our dataset from the Kaggle platform, a prominent community of data science enthusiasts. The dataset originates from the "Hotel Booking Demand Datasets" article, written by Nuno Antonio, Ana Almeida, and Luis Nunes, which was published in the Data in Brief journal in February 2019. The data cover from 06/30/2015 to 08/30/2017. The original dataset comprises 119391 rows and 32 columns, providing an extensive range of variables for analysis. Based on its size and scope, we are confident that this dataset is suitable for training an effective machine learning model.

A. Summary Analysis

In our model, we distinguish between numerical and categorical variables. The table below presents the mean and standard deviation for each numerical variable.

TABLE II
Mean and Standard Deviation of Numerical Variables

Variable	Not Canceled	Canceled
Mean		
lead_time	79.98	144.85
stays_in_weekend_nights	0.93	0.93
stays_in_week_nights	2.46	2.56
adults	1.83	1.90
children	0.10	0.11
babies	0.01	0.00
previous_cancellations	0.02	0.21
booking_changes	0.29	0.10
days_in_waiting_list	1.59	3.56
Standard Deviation		
lead_time	91.11	118.62
stays_in_weekend_nights	0.99	1.01
stays_in_week_nights	1.92	1.88
adults	0.51	0.68
children	0.39	0.41
babies	0.11	0.06
previous_cancellations	0.27	1.33
booking_changes	0.74	0.45
days_in_waiting_list	14.78	21.49

From the table above, we can see that:

- The mean lead_time for canceled bookings is significantly higher than for non-canceled bookings (144.85 vs. 79.98).
- The mean number of stays_in_week_nights for canceled bookings is slightly higher than for non-canceled bookings (2.56 vs. 2.46).
- The mean number of adults for canceled bookings is slightly higher than for non-canceled bookings (1.90 vs. 1.83).
- The mean number of previous_cancellations for canceled bookings is significantly higher than for non-canceled bookings (0.21 vs. 0.02).
- The mean number of booking_changes for canceled bookings is significantly lower than for non-canceled bookings (0.10 vs. 0.29).
- The mean days_in_waiting_list for canceled bookings is significantly higher than for non-canceled bookings (3.56 vs. 1.59).

These observations suggest that the lead_time, previous_cancellations, booking_changes, and days_in_waiting_list variables may be important predictors of cancellations in the dataset.

B. Data Cleaning

Our initial data exploration involved checking for any missing values in the dataset. Our findings indicate that the "children" column is the only one with missing data. We hypothesize that some customers who do not have children may have left this field blank, resulting in missing values. To address this, we imputed the missing values with a value of 0, which we believe is a reasonable assumption.

In addition, we observed that some of the variables in the dataset are categorical in nature and are expressed as text. For instance, the "hotel" variable denotes whether a customer has canceled their booking, the "meal" variable indicates whether a customer has made a meal reservation for the next day, and the "market_segment" variable identifies whether a customer has used a travel agent or tour operator. To process these categorical variables, we utilized dummy variables, which allow us to convert categorical data into numerical form that can be used in machine learning models.

Notably, the "reserved room type" and "assigned room type" columns indicate the type of room that the customer initially booked and the type of room they were ultimately assigned, respectively. We hypothesize that customers who are assigned a different room type than what they reserved may be more likely to cancel their reservation. To investigate this, we introduced a new column called "room_type_align", which takes on a value of 1 if the reserved room type matches the assigned room type and 0 otherwise.

After processing the variables, we randomly shuffle the dataset and divide it into training and test sets using an 80-20 ratio. To enable a more accurate comparison of the explanatory capabilities of various models, we utilize the same training and testing sets for each.

III. Analysis

In this section, we applied three different models, including logistic regression, LDA and QDA, and KNN, evaluated their performance in classifying new and unseen hotel reservations, and identified the best-performing model.

A. Logistic Regression

Logistic regression employs a sigmoid function, shown as an S-shaped curve, to map predictions and their probabilities for binary classification by converting any real value to a range between 0 and 1. Specifically,

$$f(x;\beta) = P(Y=1|X=x) = \frac{e^{\beta^T x}}{1+e^{\beta^T x}}$$
$$1 - f(x;\beta) = P(Y=0|X=x) = \frac{1}{1+e^{\beta^T x}}$$

The objective of logistic regression is to determine the optimal values of the model parameters (beta) that best fit the training data, using a technique called maximum likelihood estimation. After estimating the model parameters, the logistic regression model can be applied to make predictions on new data. If the estimated probability is higher than the predefined threshold, the model predicts that the instance belongs to that class, otherwise it predicts that it does not belong to the class.

In this study, we first standardize the input features in both the training and test sets to account for differences in scale among features. Also, this is done to enhance the performance of the optimization algorithm, such as gradient descent, which tends to converge faster and produce more reliable results when data is standardized. Next, we fit the model to the training data and learn the model parameters. These parameters are then used to make predictions on the test data, assigning labels to each record. The performance of the model is evaluated using the test error, which in this case is measured to be 0.1896, indicating that approximately 19% of the records are misclassified. Therefore, it seems that this is not the perfect method to classify the hotel booking observations, possibly due to reliance on assumptions that may not hold true for this particular dataset. And it may be necessary to develop new models that better fit the unique features of the data.

B. LDA and QDA

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are two classic classifiers with a linear and a quadratic decision surface respectively. The LDA model assumes that the covariance matrix of each class is the same, and it finds a linear decision boundary that maximizes the separation between the classes. On the other hand, the QDA model allows each class to have its own covariance matrix, and it finds a quadratic decision boundary that can fit the data more flexibly. Both LDA and QDA can be derived from simple probabilistic models which model the class conditional distribution of the data $p(\mathbf{x} \mid y = k)$ for each class k. Predictions can then be obtained by using Bayes' rule, for each training sample x,

$$p(y=k|x) = \frac{p(x|y=k)p(y=k)}{p(x)}$$

and we select the class k which maximizes this posterior probability. Bayes classifier can be rewritten as $f(x) = \arg\max_{y\in\{1,\ldots,k\}} f_y(x)$. To classify using discriminant analysis, firstly, we need to estimate k, k, k from data. Then, these estimations should be plugged into the definition of y(x) with Gaussian class conditionals. For QDA,

$$\delta_y(x) = \log(\pi_y) - \tfrac{1}{2}\log(\det(\Sigma_y)) - \tfrac{1}{2}(x-\mu_y)^T\Sigma_y^{-1}(x-\mu_y)$$

For LDA,

$$\delta_y(x) = log(\pi_y) - \tfrac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \mu_y^T\Sigma^{-1}x$$

Finally, we calculate $f(x)$, where $f(x) = \arg\max_{y\in\{1,\ldots,k\}} y_x$.

During the data-scaling process, to avoid data leakage, we computed the mean and standard deviation of the training set and used those to standardize both the training and test data. After preprocessing data, we developed and trained LDA and QDA models using the training dataset. To improve the model performance, we tried different posterior probability thresholds and selected the best threshold that maximized the accuracy value. It turned out that the optimal accuracy rate for the LDA model is 0.798 with threshold 0.44 while for QDA model is 0.667 with threshold 0.98. Therefore, the data is better suited for a linear decision boundary rather than a quadratic one. It suggests that the linear decision boundary is able to effectively separate the classes and that the data is relatively simple and well-separated. In this case, the simpler LDA model is a better choice than the more flexible QDA model, as it is less prone to overfitting and can generalize better to new data.

C. K-Nearest Neighbors

The K-Nearest Neighbors (KNN) model is designed to be a versatile and efficient classifier model. It functions by analyzing the relationships between data points within a given dataset, leveraging the spatial proximity of these points to make predictions. The model's primary principle is that data points with similar features tend to be found in close proximity to one another. As a result, KNN assigns new, unclassified data points to the class that is most common among its nearest neighbors, with the number of neighbors K. This makes the KNN model an intuitive and powerful tool for solving classification problems across various domains.

To implement the KNN model, we first identify the k nearest points to x_0 and denote them as N_0. For each distinct label, say $k = 1, 0$, we estimate the posterior probability $P(Y = k \mid X = x_0)$ as the fraction of points belonging to N_0:

$$P(Y=k \mid X=x_0) \approx \tfrac{1}{K}\textstyle\sum_{i\in N_0} \mathbb{I}(y_i = k)$$

We apply Bayes' rule and assign label k to x_0 with the largest posterior distribution. Finally, we calculate $f(x)$, where

$$\hat{f}(x) = \arg\max_{y\in\{0,1\}} P(Y=y \mid X=x)$$

We utilized the KNeighborsClassifier package from the sklearn.neighbors library to implement this process. Since the k is a very important parameter to decide the quality of the model, we tested a variety of numbers of neighbors,

specifically k=3, 5, 7, 9, 11, 13, 15, 17, and 19. Our analysis revealed that the optimal number of neighbors for maximum accuracy was k=11, with a corresponding accuracy rate of 0.8271. This demonstrates that our KNN model was successful in predicting cancellations, with an accuracy rate of 82.71%.

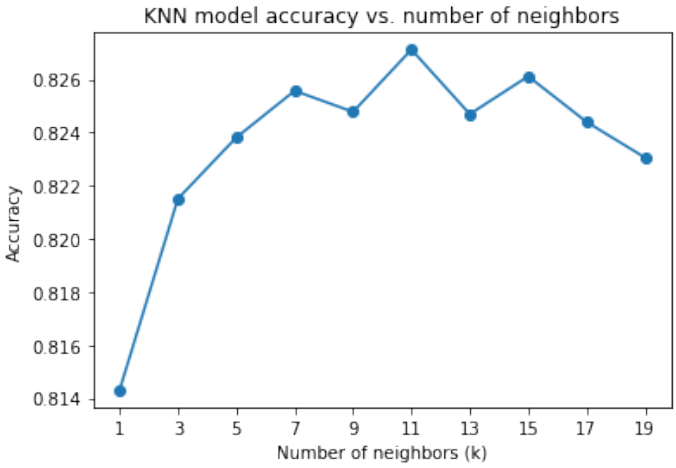

Fig. 1. KNN model accuracy vs. number of neighbors

IV. Results

After running these models on the dataset, we obtained the following accuracy rates: 81.04% for logistic regression, 79.8% for LDA, 66.7% for QDA and 82.71% for KNN. The results indicate that KNN performed the best in terms of classification accuracy. Thus, it may be a suitable choice for predicting reservation cancellations, which would enable hotel managers to make appropriate arrangements to mitigate the risks associated with cancellations.

One of the factors that could have contributed to the superior performance of KNN is the non-parametric nature, which does not assume any specific underlying data distribution. This allows KNN to capture non-linear relationships in the data, which may not be possible for logistic regression and LDA that rely on linearity assumptions. While QDA can also capture nonlinearity, it assumes different covariance matrices for each class, which may lead to poorer performance when the data has similar covariance matrices for all classes. Additionally, overfitting could be another concern with QDA as it may fit the noise in the training data. As for KNN, this would also be an issue when the value of K, the number of nearest neighbors, is small. Therefore, in addition to considering accuracy rates, it is crucial to take into account the underlying assumptions of the models when determining the most appropriate classification method.

Specifically, for logistic regression, Although the dataset meets the criteria of a large sample size and absence of extreme outliers, there may be issues with assumptions such as independence of observations, as repeated customers may place multiple orders at the hotel. Additionally, there may be multicollinearity concerns among predictor variables, as variables with multiple choices are separated into different numeric variables.

For LDA and QDA, they are both classification algorithms used to predict class labels based on input features. LDA assumes that the covariance of the predictors is equal across classes, while QDA allows different covariances. Based on the evaluation results, the LDA model performs better than the QDA model, with a higher accuracy rate. This could suggest that the LDA model has lower variance (i.e., is less prone to overfitting) but higher bias (i.e., has a simpler model structure that may not capture all the underlying complexity in the data).

The KNN model is predicated on the belief that the selected independent variables possess the capacity to accurately predict whether a customer will cancel their reservation. By examining the k nearest neighbors within the feature space, the model classifies each data point according to the predominant class among its neighboring data points. Our analysis revealed that selecting $k = 11$ neighbors yielded the highest accuracy rate of 82.71%, indicating that the chosen independent variables effectively predicted reservation cancellations. Nevertheless, it is essential to acknowledge that the model's performance is contingent on the specific dataset and feature set employed, and may exhibit variation accordingly.

V. Conclusion

After conducting extensive preprocessing and data cleaning, we developed and evaluated three classification models, including logistic regression, LDA, QDA, and KNN. Our analysis revealed that the KNN model outperformed the other models, achieving an accuracy rate of 82.71%. This suggests that KNN is the optimal model for predicting hotel room cancellations based on the information provided by the customers at the time of booking. The hotel manager can use this model to oversell rooms that have a high probability of cancellation, with the lower bound of the predicted cancellation rate serving as a guideline.

However, we still don't know which predictors have important effects on the cancellation rate. To address this, in the future analysis, we plan to use other methods, such as random forest and decision trees, to identify important predictors. This will allow us to develop specific strategies to lower the cancellation rate and improve the hotel's profitability.

Team Participation

In this project, Ivana Zhao provided an overview of the objectives of the project and our approach to addressing the room cancellation classification issue. She also conducted a literature review to examine earlier research that focused on the same topic or employed comparable machine learning techniques. In addition, she applied LDA and QDA techniques to the classification process, and afterwards, she concluded the importance of this topic after our discussion. Zhonghui Wang proposed the initial hypotheses for selecting variables, defined the project scope, and emphasized the subject matter's significance in the beginning. She then implemented the logistic regression model and compared all three models based

on their performances. And she also documented the team participation. Zeqing Zhou executed data cleaning based on variable features, and provided a comprehensive statistical summary. Additionally, she did the data analysis for the KNN model. She also gathered and refined code contributions from team members, consolidating them into the final code file. Moreover, she managed the LaTeX editing process.

To sum up, we collaborated and generated an engaging topic, discussed each step's actions, provided support and advice to each other, and together determined the best-suited model for this dataset while highlighting the importance of this topic.

REFERENCES

[1] N. Antonio, A. de Almeida and L. Nunes, "Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 2017, pp. 1049-1054, doi: 10.1109/ICMLA.2017.00-11.

[2] Falk, M. and Vieru, M. (2018), "Modeling the cancellation behavior of hotel guests", International Journal of Contemporary Hospitality Management, Vol. 30 No. 10, pp. 3100-3116. https://doi.org/10.1108/IJCHM-08-2017-0509

[3] Antonio, N., De Almeida, A. M., & Nunes, L. C. (2019). Hotel booking demand datasets. Data in Brief, 22, 41–49. https://doi.org/10.1016/j.dib.2018.11.126