DATA 1030 Midterm Project Report: Prediction Analysis of Hotel Reservation Cancelation Based on Extracted Features

Siyuan Li
Brown University
Siyuan_li2@brown.edu
GitHub-link: https://github.com/siyuanli1202/DATA1030_Midterm_Project

Abstract: Within the report, we will perform a detailed data analysis for one real multivariate dataset related to hotel bookings. The data explored contain both continuous and categorical variables. The report will extend the exploration of various classification and regression methodologies to predict the probability rate of a cancelation. The analysis also intends to enhance the prediction accuracy by reducing the data dimension with the extracted features that have the most contribution to represent the fitted model.

I. INTRODUCTION

The project is concentrated on the analysis of "*hotel_bookings.csv*" data. The resource is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, and cleaned by Thomas Mock and Antoine Bichat for research purpose. The data collected 119390 pieces of users' hotel booking information (data points), and 32 related features are recorded for each piece of information. In our analysis, we hope to use machine learning techniques to predict whether users would cancel their hotel reservations. Therefore, the target variable is a column in our dataset named "*is_canceled*". The problem can be viewed both as a regression and classification problem, as our data consists of both continuous and categorical variables. The aim for initialize the research project to solve the issue is to provide insights for hotel managers or administrators to better arrange their business operations. We would understand more clearly about the factors which affect users to retrieve their bookings. Currently, many professional projects are working on the same topic as ours, which is to predict the likelihood of a booking to be canceled. Their works shed insights on utilizing various classification models such as Random Forest and XGBoost to improve the prediction accuracy. For a detailed understanding of the features in our dataset. We summarized their attibutes in the following table.

Categorical Features *(15 counts)*	**Continuous Features *(17 counts)***
1. 'hotel' 2. 'is_canceled' 3. 'arrival_date_year' 4. 'arrival_date_month' 5. 'meal'	1. 'lead_time' 2. 'arrival_date_week_number' 3. 'arrival_date_day_of_month' 4. 'stays_in_weekend_nights' 5. 'stays_in_week_nights'

6. 'market_segment' 7. 'distribution_channel' 8. 'is_repeated_guest' 9. 'reserved_room_type' 10. 'assigned_room_type' 11. 'deposit_type' 12. 'customer_type' 13. 'required_car_parking_spaces' 14. 'total_of_special_requests' 15. 'reservation_status'	6. 'adults' 7. 'children' 8. 'babies' 9. 'previous_cancellations' 10. 'previous_bookings_not_canceled' 11. 'booking_changes' 12. 'agent' 13. 'company' 14. 'days_in_waiting_list' 15. 'adr' 16. 'required_car_parking_spaces' 17. 'total_of_special_requests'

Table 1

II. Exploratory Data Analysis

As we move to the next-step exploration of the data analysis, the key part we need to perform is EDA. To obtain a general view of the dataset, we apply the Python in-built function '*.describe or .value_counts'* to summarize the 32 features of booking information. We also create histograms and bar plots for each feature based on their attributes respectively. This statistical result is presented in the notebook. For simplification, we intended to demonstrate three interesting plots, which help us to understand the correlation of target label and rest features with a better angle. Here, we utilized three methodologies which includes bar plot, line plot and box plot. The graphs are shown as below.

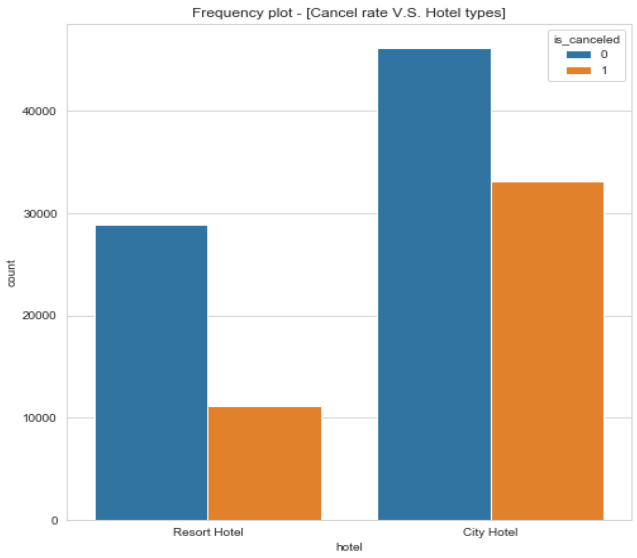

Figure 1 - Frequency plot - [Cancel rate V.S. Hotel types]

The first plot is a bar plot, which describes the relationship between booking cancelation frequency and hotel types. We can see that there is quite a visible difference between the two

hotel types: city hotel and resort hotel. The cancelation appeared more frequently at city hotels, and the reason might be caused by customer's motivation of booking. People usually book a resort hotel for spending their vacations, while more for business purposes when choosing city hotels. The difference in indentation might bring the result.

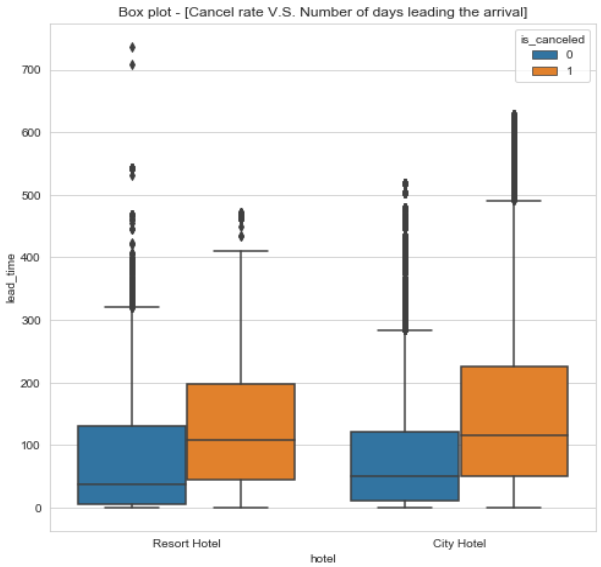

Figure 2 Box plot - [Cancel rate V.S. Number of days leading the arrival]

The second plot is a box plot, which characterized the relationship between booking cancelation frequency and leading time before the arrival data. More specifically, we plotted the correlation according to hotel types. In general, we can see that there is not an obvious difference between city hotels and resort hotels. However, the plot shows us that there is a trend appearing in both two categories. As customers' leading time increased, they are less likely to cancel their reservations. The trend might due to the fact that if people plan their trips or booking very early, they would be more determined to finish the trip and therefore take the bookings. In this way, we understand that leading time plays an important role in deciding whether a reservation would be canceled or preserved.

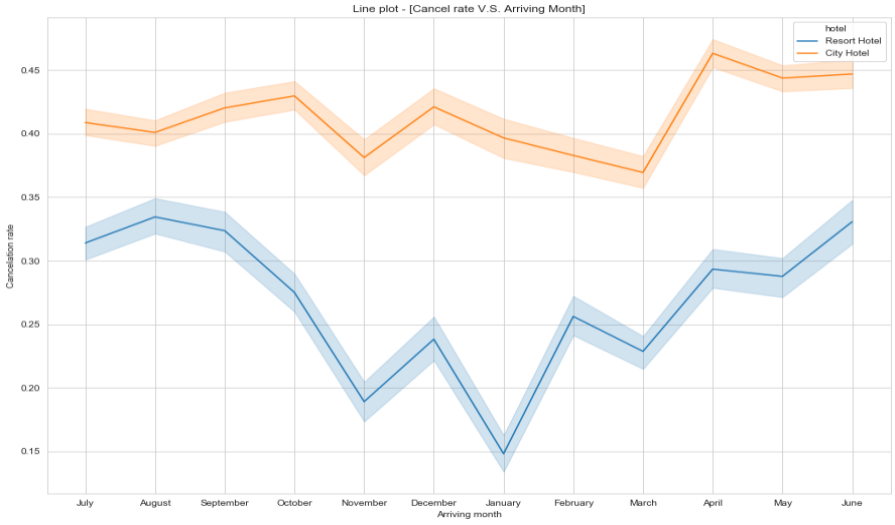

Figure 3 - "Line plot - [Cancel rate V.S. Arriving Month]"

The third plot is a line plot, which depicted the fluctuation of between booking cancelation frequency and month in a year. The plot is more designed to understand the time factor's influence on our prediction problem. With the line plot, we can discover that customers are more likely to cancel their reservation in the month like June and July, less likely in January, November and December. Accordingly, we can consider that time serves as a key factor of prediction problem.

III. DATA PREPROCESSING

As mentioned before, our prediction problem is to predict the likelihood of hotel bookings cancelation. We first clean the dataset by removing the "nan" values, which reduce a small portion of data points by 0.412%. Then we explore the dataset dimension, and find that the data frame contains about 118898 pieces of information. As the data set can be perceived as medium-sized, we will apply the traditional splitting methodology as proportion divide into 60% training, 20% validation, 20% test groups. Reflecting on the data collection source, we understand that our data is in i.i.d form. Meanwhile, it contained several time-related variables, but it is not a time series data. There are no group structures in the hotel bookings information. Given those conditions and that the data is balanced, we would directly split the dataset and apply encoders on each feature.

Now, we lead our discussion to the type of encoders utilized. Our label variable is "is_canceled", which is binary categorical variable. As the label has already been simplified into two categories, there is not necessary for now to apply "LabelEncoder" on the target. For all categorical features, we defined in the above table, we utilized "OneHotEncoder" as those features do not include ranking information. For continuous variables, we need to divide into two situations: (1) We apply "MinMaxEncoder" for two features named 'arrival_date_week_number', and 'arrival_date_day_of_month', because they are bounded in a fixed interval. (2) For the rest of continuous features, we directly use "StandardScaler". The final prepossessed training feature dataset is increasing from 29 to 1161 columns, where we havee transfomed 15 categorical features.

REFERENCES

[1] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

[2] Oliphant, T. E. (2006). *A guide to NumPy* (Vol. 1). Trelgol Publishing USA.

[3] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

[4] McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).