Data Analysis of Hotel Bookings

Group Name: JLWZ

Qiwenjing Jiang	Jialing Li	Erqian Wang	Kristine Zeng
qj336	jl9716	ew1708	yz4792

Dec 20, 2022

Introduction

During the covid-19 pandemic, the hotel industry was one of the hardest-hit industries. According to Mckinsey, the occupancy of luxury hotels was less than 15% and 40% for economy hotels in May 2020 [1]. In addition, post-pandemic travelers have changed the way they travel as well as their requirements for hotels [2]. As the main cause of low occupancy, the cancellation rate reached 40% in North America during the pandemic [3]. It is then crucial for hotels to identify potential indicators for cancellation and length of stay, which could help avoid low occupancy rates and restore profit.

Bookings Dataset

The dataset was acquired from Kaggle (Hotel Booking). It contains 119390 observations for a City Hotel and a Resort Hotel. Each observation represents a hotel booking between July 2015 and August 2017, including bookings that effectively arrived and bookings that were canceled. The meaning of each column is in appendix.

Data Preprocessing

The last few columns 'name', 'email', 'phone-number', and 'credit_card' are dropped first, since they will not be used for our analysis. We check the missing value percentage of each column and see that only four columns 'company', 'agent', 'country', and 'children' contain missing values 1. The 'children' variable indicates the number of children. As we are working on a large dataset and only 0.00335% of the 'children' column is missing, we decide to drop those rows. For the 'company' variable (the ID of the company/entity that made the booking or is responsible for paying the booking), 94.3% is missing, and we believe it is not quite significant for our analysis, we drop this column. For 'agent' and 'country', as they are categorical variables, we fill the missing values with 'unknown'. We then convert the 'hotel' column from City Hotel/Resort Hotel to 0/1, and convert 'arrival_date_month' values from month names (e.g. January) to numbers. We also add two columns 'total_nights' and 'total_people', by summing 'stays_in_weekend_nights' and 'stays_in_week_nights' of each row to get 'total_nights' for that row, and summing 'adults', 'children', and 'babies' of each row to get 'total_people'. We add these two columns since they could be important

features that generally affect the hotel booking status. Unless the booking is canceled or no-show (according to 'reservation_status' column), the rows with 'total_nights' or 'total_people' being 0 are unreasonable, so we drop these rows. In addition, we remove the outliers and negative values of the average daily rate 'adr' column. And we use one-hot encoding to convert categorical variables to dummy variables. The format of the dataset used for each question might be slightly different, so we might also further modify it in order to answer each question.

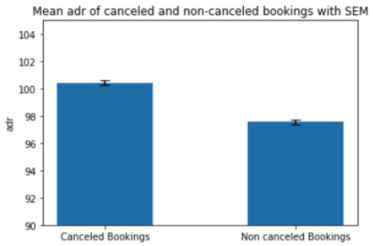
Inference Question 1

Question: Do bookings that end up getting canceled have different average daily rates than those that are not?

The question requires us to compare the average daily rate between canceled ($n = 42501$) and non-canceled bookings ($n = 71003$). To answer this question, we choose Welch's t-test to compare the average daily rates since the canceled/non-canceled bookings have different variances. In addition, we use Cohen's d to measure the effect size.
H_0: There is no difference in the daily average rate between the canceled bookings and the not canceled bookings.
H_1: There is a significant difference in the daily average rate between the canceled bookings and the not canceled bookings.

Welch's t-test

Before the Welch's t-test, we use the canceled bookings' average daily rate ($\mu_{\text{canceled}} = 100.432, \sigma^2 = 37.635$) and that of the non-canceled bookings ($\mu_{\text{non-canceled}} = 97.546, \sigma^2 = 39.379$) to determine the power of the dataset. As a result, we determined the samples have enough power ($\alpha = 0.05, 1 - \beta = 1.0$) for Welch's t-test.
The Welch's t-test indicates a significant difference ($t = 12.287, p < .001, df = 92642.088$) between the canceled bookings' average daily rate and that of the non-canceled bookings.



In addition, we split the dataset into resort hotel bookings and city hotel bookings and compare if there is a significant difference. As it turns out, the average daily rate of canceled bookings in the resort hotel is significantly different ($t = 15.661, p < .001, df = 16885.586$) from those not canceled, and the average daily rate of canceled bookings in the city hotel is significantly different ($t = -12.228, p < .001, df = 69700.224$) from those not canceled. However, for the resort hotel, the mean difference indicates a higher mean average daily rate

for the canceled bookings ($\mu_{\text{canceled}} - \mu_{\text{non-canceled}} = 8.701$) while the opposite is true for the city hotel ($\mu_{\text{canceled}} - \mu_{\text{non-canceled}} = -2.961$).

Cohen's d

In addition to significance, we also determine the effect size of the difference using Cohen's d. As a result, there is a small effect size of the average daily rate between the canceled and non-canceled bookings ($d = 0.075$). Such is true for canceled/non-canceled bookings in both the resort hotel ($d = 0.190$) and the city hotel ($d = -0.089$).

Prediction Question

Question: Does the total number of people predict total number of nights, while controlling for other explanatory variables?

Data Preprocessing for Regression

To answer the question, we need to fit a multiple linear regression, which requires more data preparation. We dropped unimportant features, alleviated skewness by log-transformation, one-hot encoded categorical features, and checked correlation matrix. For highly-correlated features, we removed one of them to avoid multicollinearity in regression. The data after preprocessing has a shape of 113504 rows and 18 columns.

Multiple Linear Regression

Because we log-transformed the response variable, the model can be expressed as

$$\log(\text{total\_nights}) = \beta_0 + \beta_1 \text{total\_people} + \cdots + \beta_{17} X_{17}$$

After splitting train and test data, we fitted a multiple linear regression using the training set. The training and testing evaluation metrics are shown in the table.

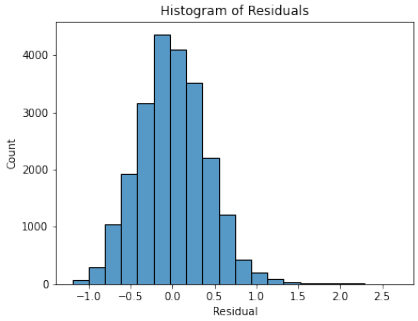
	Train	Test
R^2	0.258	0.255
RMSE	0.402	0.410

The summary of the multiple linear regression is shown below. We also plotted a histogram for residuals, showing that the residuals are centered at zero, and is roughly normal. The coefficient of determination, RMSE, and summary table demonstrate that the overall model is valid. In addition, we observed that all variables are statistically significant except several levels of certain categorical variables 2. Since we only use categorical variables as a whole, we do not remove those insignificant levels.

The explanatory variable total_people has a p-value of 5.3×10^{-63} with a coefficient estimate of 0.0447. The coefficient estimate means that keeping all other variables held constant, for every one-unit increase in total_people, we expect total_nights to increase by $\left(e^{0.0447} - 1\right) \times 100\% = 4.57\%$.

OLS Regression Results

Dep. Variable:	total_nights	R-squared:	0.258
Model:	OLS	Adj. R-squared:	0.258
Method:	Least Squares	F-statistic:	903.1
Date:	Sun, 18 Dec 2022	Prob (F-statistic):	0.00
Time:	17:08:03	Log-Likelihood:	-46195.
No. Observations:	90803	AIC:	9.246e+04
Df Residuals:	90767	BIC:	9.280e+04
Df Model:	35		
Covariance Type:	nonrobust		



Therefore, total number of people is a significant predictor for total number of nights, and one more person in the booking is expected to increase the number of nights by 4.57%.

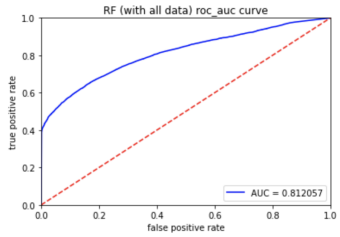
Classification Question

Question: Build a model to predict booking cancellation. Does the choice of model perform differently if we narrow down the data to certain clusters?

In order to predict cancellation, we choose to build a Random Forest classifier, and before that, we use KMeans to cluster and mark the data. We will observe and compare the Random Forest model performance on the entire dataset and on each cluster. Since clustering will be applied, we drop some columns from the dataset and only preserve the following: 'hotel', 'lead_time', 'stays_in_weekend_nights', 'stays_in_week_nights', 'adults', 'children', 'babies', 'is_repeated_guest', 'previous_cancellations', 'previous_bookings_not_canceled', 'reserved_room_type', 'deposit_type', 'total_nights', 'total_people', and the target 'is_canceled'. The numerical feature values are normalized to 0 to 1, and the categorical features are one-hot encoded. For clustering, we manually choose 5 as the number of clusters and use KMeans to form clusters based on features (target variable 'is_canceled' not included). We save the assigned cluster of each row as another feature 'cluster' for classification. The train/test is split using 0.7/0.3 proportion.

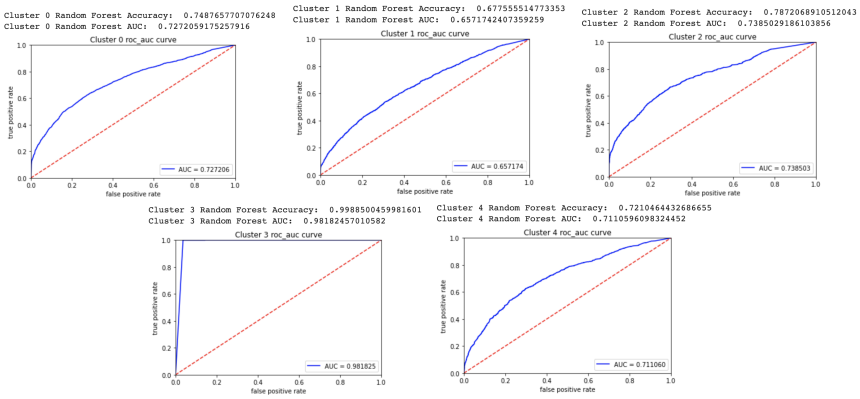
We first implement the Random Forest classifier using all train set data, use it to predict the x_test, compute its accuracy and AUC, and plot the ROC curve.

```
Random Forest Accuracy:  0.7733760131563491
Random Forest AUC:  0.8120571339370031
```



Then, we select the data of each cluster, retrain the Random Forest model for each cluster, predict the test set of the corresponding cluster, compute the accuracy and AUC, plot the

ROC curve, and show the confusion matrix correspondingly.



We can see that the overall random forest classifier achieves relatively high accuracy, 0.773, in predicting the cancellation, so 77.3% of the test set is classified correctly. The AUC, 0.812, and the ROC curve indicate that a slight increase in the false positive rate could lead to a considerable increase in the true positive rate, which means that this RF model does quite well in classifying/predicting the booking that will get canceled.

For the results of each cluster, there is no major difference among Cluster 0, 1, 2, and 4. The accuracy values are all roughly around 0.7, similar to the overall model; the AUCs are also approximately 0.7, a bit worse than the overall model. But we can see that the RF model performs the best on Cluster 3, with an accuracy of 0.99885 and AUC of 0.98182, meaning that there is minimal error while classifying the cancellation result of Cluster 3. We observe Cluster 3 and see that the bookings mostly get canceled in the end (true value), and these bookings are mostly made by 2 adults without children or babies, the lead_time is long, and the bookings are usually for just one or two nights during the weekdays. Also, those people mostly are not repeated guests, and the reserved room type is mostly A (probably the small ones). From these data, we can infer that Cluster 3 is possibly young people who are more flexible in time and plans, so they are more likely to cancel their bookings, and thus the RF model does the best in predicting/classifying this group of people.

In addition, we notice that the feature importance for lead_time variable is higher than other features in all these RF models 3, so we further analyze the lead time and its relation to cancellation in the following section.

Extra Credit: Inference Question

Question: Is the lead time for users who canceled their reservations significantly greater than that for users who did not?

Lead time is defined as the number of days elapsed between the entering date of the booking into the PMS and the estimated arrival date. The first guess we had is that the lead time for users who canceled their reservations is greater than that for users who did not. This is

because people who made their reservations long before their estimated arrival date may be more likely to have their plans changed due to unpredictable circumstances. To test if the guess is valid, we formulated the following statement and hypotheses.

Statement: the lead time for users who canceled their reservations is significantly greater than that of users who did not.

Let X_i denote each data point from the cancellation group; let Y_j denote each data point from the non-cancellation group. Then our hypotheses are written as

$$H_0 : P(X_i > Y_j) \leq 0.5$$
$$H_1 : P(X_i < Y_j) > 0.5$$

Characteristics of sample groups: our two groups are independent. They have different sizes, variances, means, and similar distribution shapes. Data are continuous.

Power Analysis, Mann-Whitney U Test and Effect Size

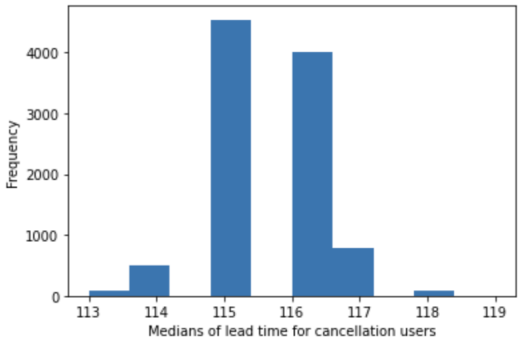
Since our sample size exceeds 20000, which is large enough to detect the effect at the desired significance level, we can proceed further to do the test. Based on the above characteristics, we chose the non-parametric test, specifically, the Mann-Whitney U test. The test statistic $U = 2078110730.5$ and the corresponding p-value is smaller than 0.00005. Both values are in their reasonable ranges. By comparing with the critical p-value of 0.05, we conclude with a 95% confidence that the lead time is significantly greater for users canceling their reservation. In other words, the probability of each X_i being greater than each Y_j is bigger than 0.5, and we reject the null hypothesis.

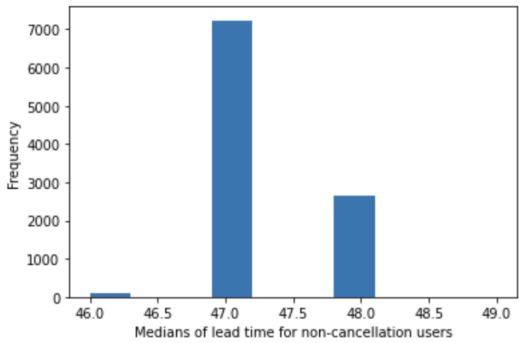
However, a test of statistical significance is only a test of the plausibility of the model represented by the null hypothesis. Therefore, the Mann-Whitney U test cannot tell us how important a result is. In order to interpret the meaning of the above result, we need to calculate the effect size, and we get $r = 0.68$, which is greater than 0.5, meaning the effect is large.

Through bootstrapping, each bar plot below shows 10000 instances of medians sampled from the original data set.

Conclusion

Thanks to the many tools provided by DSGA-1001, we are able to dive deeper into the hotel booking dataset and get our questions answered.
Takeaway: We identified a significant difference in average daily rate between canceled and non-canceled bookings with limited effect size, which is valid for both resort and city hotel bookings. In addition, we found lead time for cancellation users is significantly greater than that for non-cancellation users. We built a valid multiple linear regression to predict the length of stay, and we have evidence to state that the number of people is a significant predictor of the number of nights. Also, we found that the random forest can serve as a valid






```
95.0% confidence interval is between 114.0 and 117.0
```

```
95.0% confidence interval is between 47.0 and 48.0
```

classifier for predicting cancellation as it achieves high accuracy, so we can apply it to predict future hotel bookings. In addition, we noticed, through comparing RF model performance across different clusters, a group of people that can be more easily predicted. They are likely to be young adults without kids who are more flexible in time and plans, and thus they tend to cancel hotel bookings and can be predicted more accurately. So for future bookings, if they are marked to this cluster, the RF model specific to this cluster can be used to predict their cancellation.

Assumptions: The t-test requires the data to be normal. The Mann-Whitney U test requires data to be continuous and independent; the samples should also have similar distributions. For multiple linear regression, we assumed that there is a linear relationship between each predictor and the outcome and that the residuals are independent and follow a normal distribution of mean zero and a constant variance.

Limitations: In the beginning, we mentioned that we are interested in avoiding loss for hotels post-pandemic. However, the data is between 2015 and 2017. Also, the data only include bookings of two hotels, so it may not be generalized well for other hotels. Statistically, we assumed that the explanatory variables are independent and normal. However, a number of them are skewed. Ideally, the data set is post-pandemic and contains information of a variety of hotels. For example, the data set is collected from a large booking website like Expedia. Also, the features should independently follow normal distributions.

Author Contributions

Qiwenjing Jiang: introduction, inference question, conclusion
Jialing Li: inference question for extra credit, conclusion
Erqian Wang: prediction question, conclusion
Kristine Zeng: data preprocessing, classification question, conclusion

References

[1] Mckinsey

[2] The Seattle Time

[3] Knowland

Appendix

Column Details

1. hotel: The type of hotel being booked (resort/city).
2. is_canceled: 1 if the booking was canceled, 0 if not.
3. lead_time: Number of days between the entering date of the booking and the arrival date.
4. arrival_date_year: Year of arrival date.
5. arrival_date_month: Month of arrival date with 12 categories: "January" to "December".
6. arrival_date_week_number: Week number of the arrival date.
7. arrival_date_day_of_month: Day of the month of the arrival date.
8. stays_in_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
9. stays_in_week_nights: Number of weekday nights (Monday to Friday) the guest stayed or booked to stay at the hotel.
10. adults: Number of adults
11. children: Number of children
12. babies: Number of babies
13. meal: Bed & Breakfast
14. country: guest's country of origin
15. market_segment: Market segment designation (TA: Travel Agents, TO: Tour Operators).
16. distribution_channel: Booking distribution channel (TA: Travel Agents, TO: Tour Operators).
17. is_repeated_guest: 1 if the booking is from a repeated guest, 0 if not.
18. previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking.
19. previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking.
20. reserved_room_type: Code of room type reserved.
21. assigned_room_type: Code for the type of room assigned to the booking.
22. booking_changes: Number of changes/amendments made to the booking from the moment of booking until the moment of check-in or cancellation
23. deposit_type: The deposit type of the booking(No Deposit/Non Refund/Refundable)
24. agent: ID of the travel agency that made the booking
25. company: ID of the company/entity that made the booking or is responsible for paying the booking. ID is presented instead of designation for anonymity reasons
26. days_in_waiting_list: Number of days the booking was in the waiting list before it was confirmed to the customer
27. customer_type: The type of customer(Group/Transient/Transient-party)
28. adr: Average Daily Rate (Calculated by dividing the sum of all lodging transactions by the total number of staying nights)
29. required_car_parking_spaces: Number of car parking spaces required by the customer
30. total_of_special_requests: Number of special requests made by the customer (e.g. twin bed or high floor)
31. reservation_status: The status of booing(Check-out/No-Show)
32. reservation_status_date: Date at which the reservation status was set.
33. name: Name of the Guest (not real)
34. email: Email (not real)
35. phone-number: Phone number (not real)
36. credit_card: Credit Card Number (not real)

	missing value percentage
company	94.306893
agent	13.686238
country	0.408744
children	0.003350
reserved_room_type	0.000000
assigned_room_type	0.000000
booking_changes	0.000000
deposit_type	0.000000
hotel	0.000000
previous_cancellations	0.000000
days_in_waiting_list	0.000000
customer_type	0.000000
adr	0.000000
required_car_parking_spaces	0.000000
total_of_special_requests	0.000000
reservation_status	0.000000
previous_bookings_not_canceled	0.000000
is_repeated_guest	0.000000
is_canceled	0.000000
distribution_channel	0.000000
market_segment	0.000000
meal	0.000000
babies	0.000000
adults	0.000000
stays_in_week_nights	0.000000
stays_in_weekend_nights	0.000000
arrival_date_day_of_month	0.000000
arrival_date_week_number	0.000000
arrival_date_month	0.000000
arrival_date_year	0.000000
lead_time	0.000000
reservation_status_date	0.000000

Figure 1: Missing Value Percentage

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.3009	0.031	42.224	0.000	1.241	1.361
hotel	0.2297	0.004	63.462	0.000	0.223	0.237
is_canceled	0.0202	0.003	5.780	0.000	0.013	0.027
lead_time	0.1142	0.001	112.037	0.000	0.112	0.116
is_repeated_guest	-0.0743	0.009	-8.329	0.000	-0.092	-0.057
previous_cancellations	-0.0718	0.007	-9.914	0.000	-0.086	-0.058
booking_changes	0.0700	0.004	15.962	0.000	0.061	0.079
days_in_waiting_list	-0.0102	0.002	-5.296	0.000	-0.014	-0.006
adr	0.0004	4.32e-05	9.914	0.000	0.000	0.001
required_car_parking_spaces	-0.0943	0.006	-15.873	0.000	-0.106	-0.083
total_of_special_requests	0.0120	0.002	6.069	0.000	0.008	0.016
total_people	0.0447	0.003	16.767	0.000	0.039	0.050
meal_FB	0.0024	0.018	0.134	0.893	-0.032	0.037
meal_HB	0.0122	0.005	2.660	0.008	0.003	0.021
meal_SC	-0.0305	0.005	-5.911	0.000	-0.041	-0.020
meal_Undefined	0.0993	0.014	6.941	0.000	0.071	0.127
market_segment_Complementary	-0.4027	0.065	-6.165	0.000	-0.531	-0.275
market_segment_Corporate	-0.4753	0.030	-15.752	0.000	-0.534	-0.416
market_segment_Direct	-0.3586	0.030	-11.979	0.000	-0.417	-0.300
market_segment_Groups	-0.4437	0.030	-14.777	0.000	-0.503	-0.385
market_segment_Offline TA/TO	-0.2942	0.030	-9.840	0.000	-0.353	-0.236
market_segment_Online TA	-0.3266	0.030	-10.957	0.000	-0.385	-0.268
assigned_room_type_B	-0.0389	0.010	-3.852	0.000	-0.059	-0.019
assigned_room_type_C	-0.0576	0.011	-5.460	0.000	-0.078	-0.037
assigned_room_type_D	0.0262	0.004	7.225	0.000	0.019	0.033
assigned_room_type_E	0.0466	0.006	7.770	0.000	0.035	0.058
assigned_room_type_F	-0.0430	0.009	-4.736	0.000	-0.061	-0.025
assigned_room_type_G	-0.0477	0.012	-3.984	0.000	-0.071	-0.024
assigned_room_type_H	-0.1563	0.021	-7.363	0.000	-0.198	-0.115
assigned_room_type_I	-0.0117	0.036	-0.321	0.748	-0.083	0.060
assigned_room_type_K	-0.0599	0.040	-1.492	0.136	-0.139	0.019
assigned_room_type_L	-0.3116	0.403	-0.774	0.439	-1.101	0.477
deposit_type_Non Refund	-0.1654	0.006	-29.951	0.000	-0.176	-0.155
deposit_type_Refundable	-0.0251	0.036	-0.705	0.481	-0.095	0.045
customer_type_Group	-0.2917	0.022	-13.551	0.000	-0.334	-0.249
customer_type_Transient	-0.2304	0.007	-31.161	0.000	-0.245	-0.216

Figure 2: Multiple Linear Regression Summary

feacture	overall RF importance	Cluster 0 importance	Cluster 1 importance	Cluster 2 importance	Cluster 3 importance	Cluster 4 importance
hotel	0.011060	0.000000	0.016151	0.000000	1.032643e-01	0.017921
lead_time	0.500399	0.752035	0.800441	0.765914	5.071738e-01	0.635289
stays_in_weekend_nights	0.016604	0.017784	0.022468	0.025217	8.927895e-02	0.038328
stays_in_week_nights	0.026674	0.029143	0.036104	0.040495	6.622537e-02	0.061682
adults	0.007948	0.011678	0.011840	0.012305	1.931195e-02	0.018293
children	0.005966	0.004946	0.004433	0.003836	8.762302e-03	0.028406
babies	0.001368	0.001005	0.001555	0.002015	0.000000e+00	0.004529
is_repeated_guest	0.003700	0.008794	0.003178	0.005682	6.353448e-03	0.003402
previous_cancellations	0.060812	0.117455	0.022784	0.063546	2.347344e-02	0.027366
previous_bookings_not_canceled	0.010103	0.014856	0.005793	0.017716	2.621013e-02	0.007389
total_nights	0.031229	0.031499	0.040263	0.047368	9.848628e-02	0.066799
total_people	0.010251	0.010210	0.010065	0.013332	1.668024e-02	0.031138
reserved_room_type_A	0.006015	0.000000	0.000000	0.000000	1.285002e-02	0.000000
reserved_room_type_B	0.001023	0.000000	0.000000	0.000000	2.506677e-03	0.005951
reserved_room_type_C	0.000945	0.000000	0.000000	0.000000	3.916236e-07	0.006610
reserved_room_type_D	0.002076	0.000000	0.000000	0.000000	7.156988e-04	0.000000
reserved_room_type_E	0.003279	0.000000	0.000000	0.000000	1.807278e-02	0.011310
reserved_room_type_F	0.001735	0.000000	0.000000	0.000000	0.000000e+00	0.009079
reserved_room_type_G	0.001328	0.000000	0.000000	0.000000	0.000000e+00	0.008560
reserved_room_type_H	0.000647	0.000000	0.000000	0.000000	6.341878e-04	0.004727
reserved_room_type_L	0.000017	0.000000	0.000000	0.000000	0.000000e+00	0.000119
deposit_type_No Deposit	0.102020	0.000317	0.013571	0.001274	0.000000e+00	0.005321
deposit_type_Non Refund	0.144392	0.000000	0.011154	0.000000	1.709405e-08	0.007184
deposit_type_Refundable	0.000982	0.000278	0.000200	0.001299	2.735699e-08	0.000596

Figure 3: Random Forest Classifiers Feature Importances