Hotel Booking Demand Analysis: Customer Segmentation and Cancelation Prediction

Abstract

In this project, data analysis is conducted on a real-world dataset named *Hotel Bookings Demand* to investigate patterns on hotel booking behaviors. The dataset is the booking records provided by two hotels, one located in city and the other located in resort, in Portugal from 2015 to 2017. It is of great importance for hotel managing purpose to figures out the answers to following two questions: (i) How to predict whether a customer would cancel the booking or not? (ii) Are there any patterns in customers' booking behaviors, by which they can be segmented into groups that enable the managing team to propose distinct promotions? Hence, customer segmentation and cancelation prediction are investigated. Customer segment is done by dimension reduction via t-SNE, followed by hierarchical clustering. Cancelation prediction is done by applying Random Forest Classification. It is found that there are 7 customer groups with distinct booking patterns. The cancelation prediction has achieved an accuracy of 80% and 79% for train and test set, respectively.

1. Introduction

Booking demand is the essential part for hotel business for both customers and hotel managers. A smart booking system should not only be able to satisfy customers' requests, but more importantly, enable hotel runners to effectively manage their resources and rapidly response to customer requests. A promising solution is to introduce machine learning algorithms into a booking system, commercial products being available now.[1] Customer segmentation and cancelation prediction are two topics drawing most attentions. Customer segmentation is to group booking records or customer information into a few clusters so that each cluster shares characteristics different from other clusters. Customer segmentation essentially provides indispensable insights on understanding the underlying structure of customers, which enables hotel runners to adjust their strategies on sending promotions or simply serves as a new predictor in a more complicated machine learning model. A promising way for customer segmentation is to use unsupervised machine learning algorithms like clustering. To reduce the training time, the dataset is usually processed by dimension reduction algorithms like PCA and t-SNE before clustering. Cancelation Prediction, as the name indicates, is to predict if a booking will be canceled. It is helpful to know the percentage of "fake" bookings especially in the high-demand periods such as statutory holidays. Most classification algorithms should work, but, considering the outliers and skewness of the hotel booking dataset that will be discussed later in this report, a tree-based classifier like random forest would be a perfect candidate.

2. Data Preparation

2.1. Metadata

Hotel Booking Demand dataset is first introduced by N. Antonio and coworkers in 2019, [2] and is accessible on Kaggle.com.[3] The metadata is shown below.

Table 1. Metadata of Hotel Booking Demand dataset

Number of Rows	119390
Number of Columns	32
Number of Numerical Variables	17
Number of Categorical Variables	13
Number of Date Variables	1

2.2. Duplicate Removal

As the identification information of customers and hotels are removed from the dataset for privacy purpose, there are duplicate booking records in the dataset. 31994 out of 119390 records are found duplicate. To prevent data leakage and avoid adding unexpected weights in modeling, the duplicate records are removed, keeping the counts as a new columns named *count_of_record*. The dimension of dataset is now 87396 rows and 33 columns.

2.3. Missing Value Imputation

All non-zero percentages of missing values are shown in the table below, as well as the imputed vales as advised in the source paper.

Table 2. Missing percentages and imputed values

	children	**country**	**agent**	**company**
Missing Percentage	0.0046 %	0.52 %	14 %	94 %
Imputed Value	0	UNK	NA	0

UNK in *country* is short for unknown; NA in *agent* is short for not applicant; and 0 in *company* indicates the booking is paid by a non-company entity, or by an individual.

2.4. Transform Columns

Company: 353 unique values in this categorical feature. As 94% of records are marked 0 in company, the 6% non-zero values are converted to 1 uniformly. Company is now binary.

Country: 178 unique values in this categorical feature. The top 3 frequent country labels are kept as they are, the others are converted into OTHER, resulting in 4 unique values.

Agent: 334 unique values in this categorical feature. The top 3 frequent agent labels are kept as they are (NA is one of them), the others are converted into OTHER, resulting in 4 unique values.

Arrival date month: Names of month are converted into ordinal values ranging from 1 to 12.

Reservation status date: The data type of date column is converted from string to datetime built in Python.

2.5. Outlier Removal

The distributions of numerical and ordinal variables are found highly skewed with long right tails, including continuous variables, like average daily rate and lead time, and concrete variables, like number of adults, children and babies. Outlier detection using z-score is not an option for such distributions. Despite applying logarithm function to reduce skewness may work, a more robust way is to select algorithms insensitive to feature distributions, such as random forest. Hence, thresholds are manually assigned to remove outliers based on domain knowledge.

Table 3. Thresholds for outliers

Average Daily Rate	Lead Time	Days in Waiting List
< 1000	< 700	< 300

3. Exploratory Data Analysis

3.1. Numerical Variables

Distributions of most numerical variables in the datset are highly skewed. It is easy to understand that they tend to show a value closed to zero. For example, number of adults is usually 1 or 2 rather than over 10; booking changes are usually 0 rather than a higher number. For such distributions, it is problematic to identify outliers with z-score > 3. Another way is to transform distributions to be normal-distribution-like using logarithm function y=lg(x+1), but, for highly skewed features like previous cancelation or number of children, it is still not good enough. Therefore, linear-based models, like linear regression and logistic regression, are not good candidates, while tree-based models are proper for cancelation prediction, due to their insensitivity of outliers and distribution.

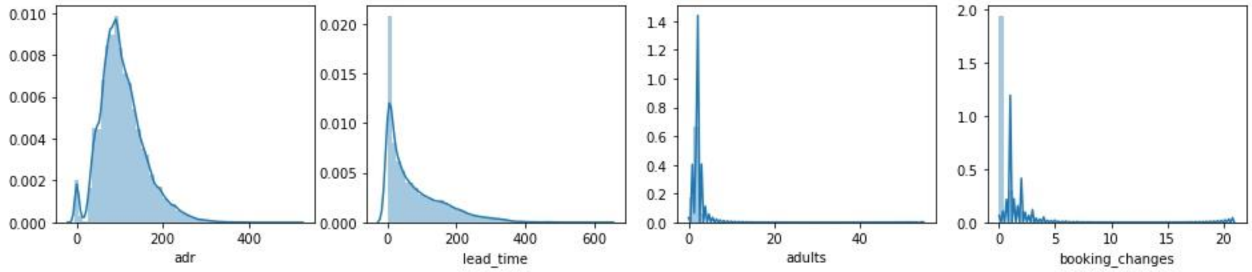

Figure 1. Distributions of a subset of numerical variables.

Correlation Coefficients for each numerical variable pair is calculated and shown in Figure 2. The variable pairs with high correlation coefficients will be further investigated later.

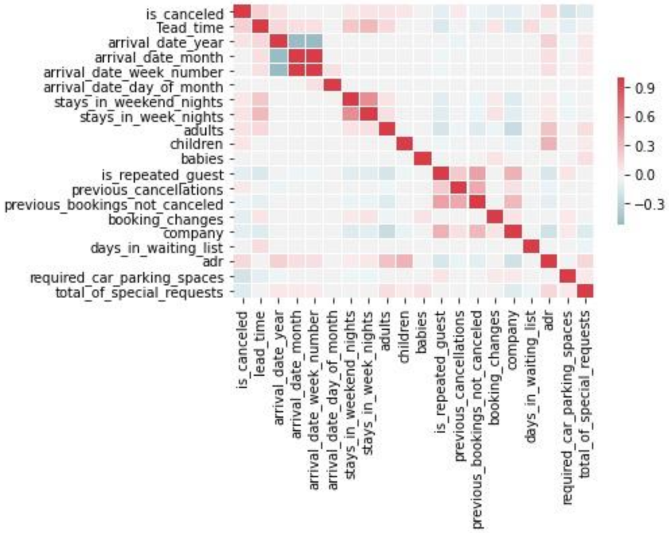

Figure 2. Heat map of Pearson's Correlation for numerical variables.

3.2. Categorical Variables

Cancelation Ratio is a critical variable in the project, so it is essential to explore what variables enable us to predict cancelation.

Booking Demand by month is shown in Figure 3. Cancelation ratios in high-demand months are higher compared to the other months.

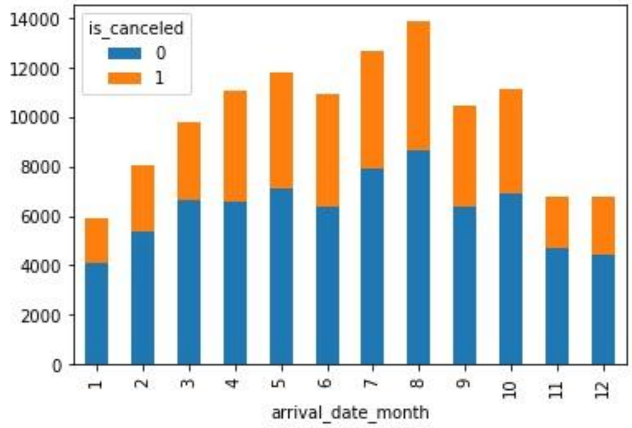

Figure 3. Booking demand and cancelation by month.

Customer-Related Variables plays significant roles in cancelation. As shown in Figure 4, repeated guests are less likely to cancel a booking. Transient customers are more likely to cancel a booking. More interestingly, bookings with less special requests are more likely to be canceled.

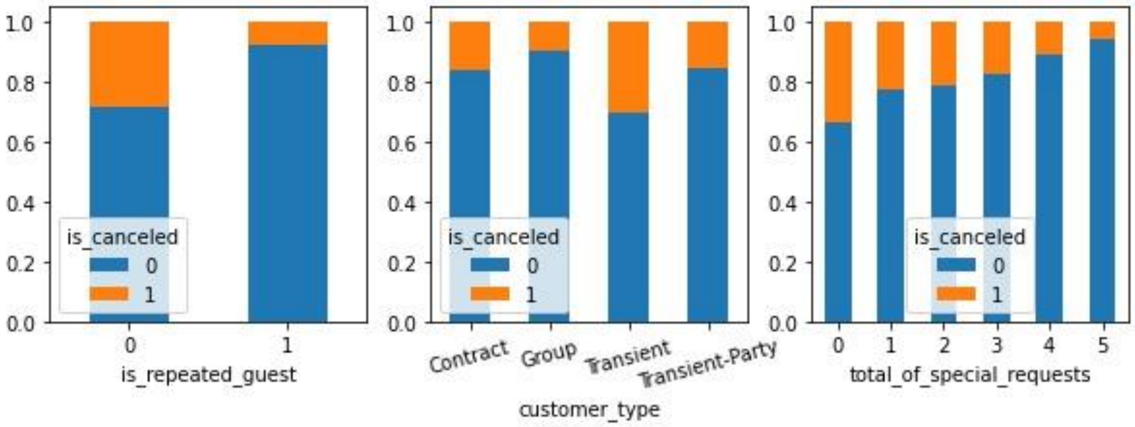

Figure 4. Cancelation dependent on customer-related variables: *is repeated guest*, *customer type* and *total of special requests*.

Room-Related Variables are also related to cancelation. By comparing *reserved room type* and actually *assigned room type*, a new variable is created and added to the dataset named *changed room type*, which has binary values, True if reserved and assigned room type are the same, and False if not. It can be observed in Figure 5 that bookings with room type changes are less likely to be canceled. Besides, room types can be divided into 3 groups: I and K are of lowest cancelation ratio; P are of highest cancelation ratio; the rest (A-L) shares similar medium cancelation ratios.

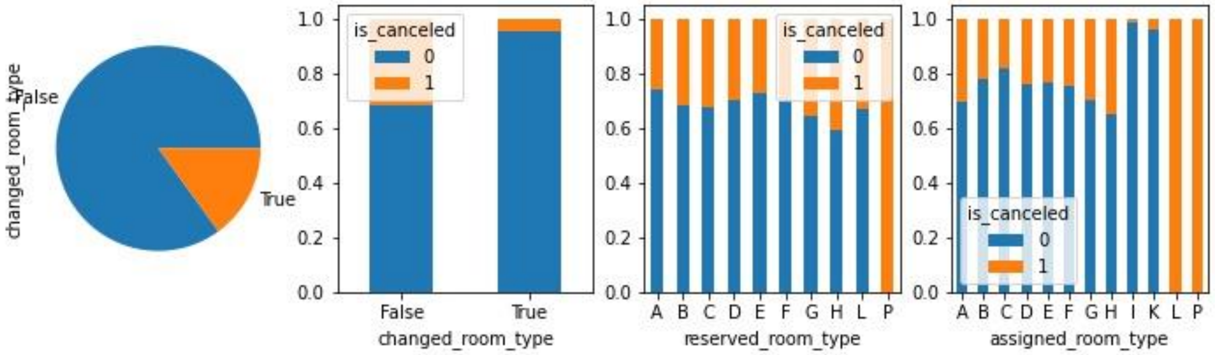

Figure 5. Cancelation dependent on room-related variables: *changed room type*, *reserved room type* and *assigned room type*.

4. Data Processing

According to the results in EDA, a new feature named *changed room type* is added, while *count of record* and *reservation status date* are dropped because they are not related to either customer segmentation or cancelation prediction.

5. Customer Segmentation

5.1. Feature Selection

Customer segmentation will include clustering. Distance-based clustering algorithms are significantly limited by curse of dimension, so features should be carefully selected for clustering. There are 13 categorical features in the dataset till this step. As one-hot-encoded categorical features will lead to overwhelmingly high-dimensional feature space, only binary and ordinal categorical features will be used for customer segmentation. Numerical features will not cause an increase in dimension, so all can be used for clustering. Hence, 20 features in total are selected for customer segmentation.

5.2. Dimension Reduction

Principle Component Analysis (PCA) is the most widely used unsupervised model for dimension reduction. However, Elbow Method failed in the dataset when trying to find an optimal number of principle components, as shown in Figure 6(a). Compared to PCA, t-SNE, as another dimension reduction algorithm, is popular for maintaining the microstructures while reducing dimension, which is perfect for visualization. To better understand the structure of dataset in feature space, t-SNE is applied to reduce feature space to 2-dimensional, which is shown in Figure 6(b). The two scatter plots in Figure 6(b) are identical except for the coloring (to be discussed in Section 5.3). Perplexity and learning rate are two hyper parameters dominates the performance of t-SNE. A perplexity of 30 and a learning rate of 300 are found optimal in showing separated clusters.

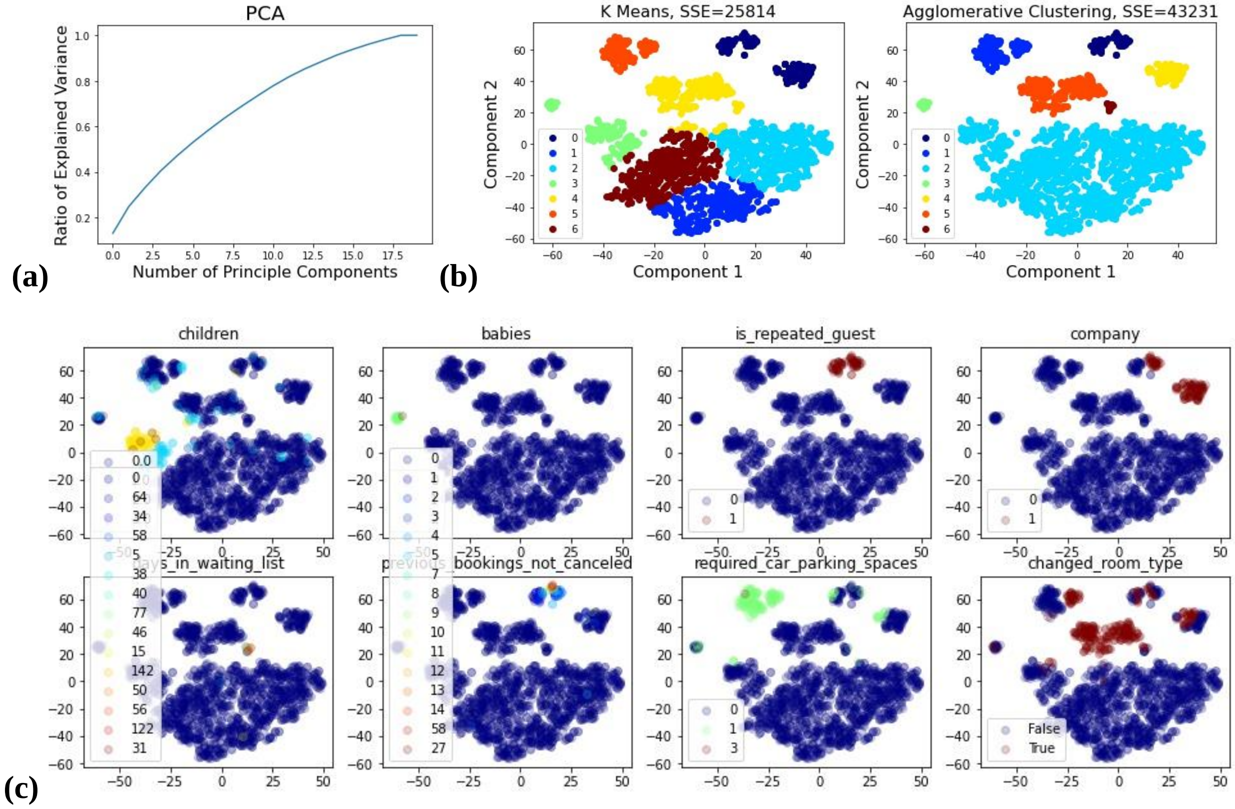

Figure 6. (a) Explained variance vs number of principle components in PCA. (b) 2-dimensional scatter plot generated by t-SNE algorithm and clustered by K Means and Agglomerative Clustering, respectively. (c) Coloring data points by categories for 8 different features used in clustering.

5.3. Clustering

As the 2-D data generated by t-SNE are basically convex and sphere in shape, K Means and Agglomerative Clustering are promising to achieve good enough clustering performance. It can be recognized that there are one large clusters in the bottom and 6 small ones on top on it, so it is reasonable to consider K = 7. The clustering results of K Means and Agglomerative Clustering are shown in Figure 6(b), colored by cluster label. The metric to evaluate clustering is Sum of Squared Errors (SSE). K Means obviously achieved higher SSE and uniform-size clusters, but Cluster 1, 2 and 6 are not quite separable. Clusters generated by Agglomerative Clustering are more closed to what is expected, despite of higher SSE.

To further evaluate the results of the two models, the 2-D dataset is colored by categories extracted from each one among 8 important features found in EDA: number of children, number of babies, whether the guest is a repeated guest, whether the booking is made by a company, days in waiting list, previous bookings not canceled, required car parking spaces, and changed room type. The corresponding figures are shown in Figure 6(c). Among the 8 clusters recognizable in the figures, K Means recognizes 4, while Agglomerative Clustering recognizes 7 (fails only on the children cluster). Therefore, Agglomerative Clustering is a better candidate for the 2-D feature space.

5.4. Discussion

Select Optimal K Value. When assign K=8 or 9, none of the two algorithms fail to recognize the children cluster. Instead, more noise clusters are generated, so K=7 is the optimal value.

Scalability. A drawback of the model is the scalability. As t-SNE are computationally expensive and Agglomerative Cluster requires O(n^2) space complexity, the model was tried only on pilot datasets, but failed to run on the whole dataset (80k+ rows) due to limited memory. To scale the model to handle large datasets, Agglomerative Clustering should be avoided.

Profile Customer Behavior Pattern. Combining all the results in Figure 6, we are able to roughly profile the behavior patterns of each customer segment, as summarized in Table 4.

Table 4. Characteristics of each customer cluster generated by Agglomerative Clustering.

Customer Cluster #	Characteristics
0	previous_booking_not_canceled > 0; is_repeated_guest == True
1	required_car_parking_spaces > 0
2	all bookings excluded by other clusters
3	babies > 0
4	company == 1
5	changed_room_type == True
6	days_in_waiting_list > 0

6. Cancelation Prediction

The dataset is split to train set and test set with a ratio of 2:1. The target variable is *is canceled*. Ordinal and binary variables are grouped together with numerical variables to be numerical features, which will be processed with standard scaler. Categorical features are one-hot-encoded. Dimension of features space is now 77.

6.1. Model Selection

Due to the skewness of the dataset, tree-based classification models are a promising candidate. Therefore, Random Forest is selected for cancelation prediction. To select an optimal model, hyperparameters are tuned by grid search together with Stratified K Fold cross validation. Due to the imbalanced classes in target, hyperparameter class_weight is assigned “balanced”.

Table 5. Features and corresponding parameters for grid search.

max_depth	max_features	n_estimators	min_samples_leaf	min_samples_split
5, 10	0.5, sqrt	50, 100, 200	2, 5	5, 10

The best model is {'max_depth': 10, 'max_features': 0.5, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 50}.

6.2. Performance Evaluation

The performance of classification is shown in Table 6.

Table 6. Performance of classification model

Train Set	Test Set

is_canceled	Precision	Recall	Accuracy	is_canceled	Precision	Recall	Accuracy
0	0.94	0.77	0.80	0	0.93	0.77	0.79
1	0.59	0.87		1	0.58	0.85	

Despite assigned balanced weight to suppress imbalanced classes in target, minority class 1 still performs worse than class 0. Performance between train and test set is not significant. The model is high- bias while low-variance.

6.3. Discussion

Class Imbalance should be handled in a better way. To increase the precision of minority class, an effective way is to down-sampling the majority to be of the same size as minority. Another way is to up-sampling the minority by synthesizing minority class observations using SMOTE.

Feature Importance is calculated while training the model. As shown in Figure 7, the two most important features are lead time (how long the booking is made before arrival) and country PRT (whether the customer is local in Portugal). Early bookings are more likely to be canceled. Foreign customers are more cautious on canceling bookings when traveling in another country.

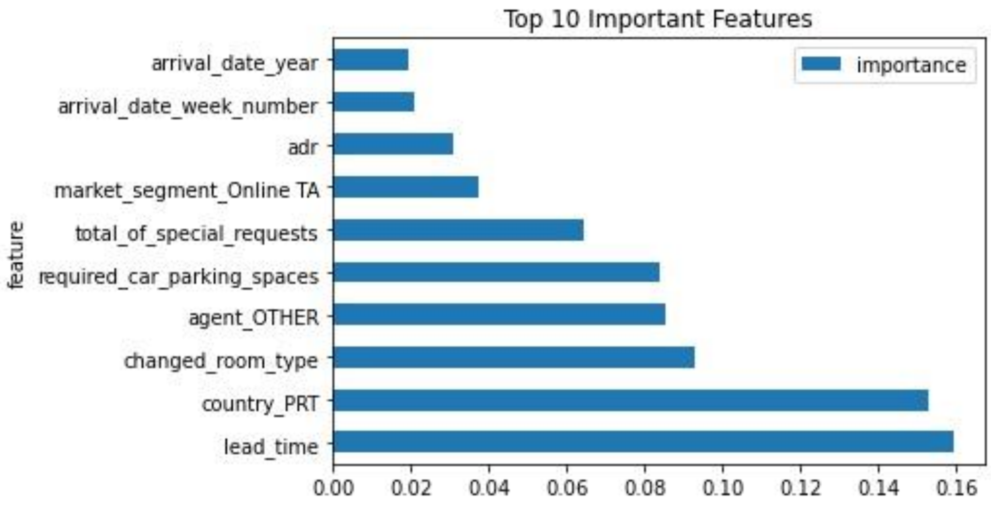

Figure 7. Top 10 important features.

7. Conclusion

Customer segmentation and cancelation prediction are conducted on the dataset of hotel booking records. 7 customer groups are recognized and profiled by customer segmentation. The accuracy of cancelation prediction achieved 80% and 79% for train set and test set, respectively.

References

1. Allora AI Booking Platform, https://www.allora.ai/
2. Hotel booking demand datasets, Data in Brief, Vol 22 (2019) 41-49
3. Hotel booking demand datasets on Kaggle.com, https://www.kaggle.com/jessemostipak/hotel-booking-demand