|  |  |
| --- | --- |
| LIN Yu-Hui  LIU Xinran  TSAO Yu-Lun | AN Ning Pooja Nitin WADHWA  SHI Yawen |

Delivering Customer-Centric Hotel Booking Service through Customer Analytics

*Abstract*

In the current data-rich environment, more and more businesses make proactive attempts to understand their customers better through the use of customer analytics to further bring their service to a higher level. In this report, we place our focus on hotel booking data provided by Expedia, a leading online platform that is widely recognized for travel booking, to explore how the business can better utilize their customer data to more effectively acquire and develop customers through the customer lifecycle.

# Introduction and Motivation

As the travel industry continues to evolve, it is critical to analyze related hotel booking data to not only understand the current trends and customer preferences, but also identify future opportunities in terms of developing and expanding customer base.

Expedia, as an online booking platform, with over 600 million monthly users on its site and accommodations in more than 200 countries (Heng, 2019), is likely to be interested in retaining customers and hotel merchants. As a customer centric organisation, they are motivated to deliver customized service offerings to their customers.

Our analysis is designed to leverage customer data to allow Expedia to offer customized services to its customers. To achieve this customization of services offered on the site, we have structured our analysis in 4 main segments with reference to the first stages of the customer lifecycle (Linoff & Berry, 2000), namely, predicting the probability of booking, customer segmentation with RFM analysis, market basket analysis and finally designing a recommendation system based on customer behaviors.

In the subsequent sections, we will start by contextualizing our analysis with details on the dataset used and initial analysis through exploratory data analysis. Analytical techniques, insights and findings will be further discussed for each of the four analysis we have performed and finally we will wrap up with the conclusion and evaluations of the analysis.

# Dataset

The chosen dataset is sourced from Kaggle.com and provided by the Expedia Group. The dataset includes from January 2013 to January 2015. The data provides detailed logs of customer behaviors when they visit the site including customers’ response to the search results (click/book) and the conditions they set when searching in the site. Additionally, the data includes details such as the customers’ geographical location, features collected by the platform like the marketing channel used, the existence of a travel package, and so on.

Most features in the data are numeric and the data attributes included in the dataset are listed in detail in Appendix 1.

Our dataset does not explicit include the uniquely identified hotels offered by Expedia.com, however, the dataset does include an attribute named ‘Hotel \_cluster’, which refers to hotel clusters derived with Expedia’s in-house algorithms, where similar hotels for a search (based on historical price, customer star ratings, geographical locations relative to city center, etc) are grouped together. These hotel clusters could be used as identifiers to hotel types.

Another caveat of the dataset is that all the location-related data are encoded in integer forms as compared to textual character form as how geographical locations are generally identified. Given that the dataset is provided by Expedia.com, for confidentiality purposes, the specific location of the hotels and customers are not disclosed such that our analysis would not be able to explicitly distinguish the geographical locations.

# Exploratory Data Analysis

For the purpose of this analysis, data from January 2013 to January 2015 has been considered. Data cleaning steps such as dropping duplicate records, observations with missing key values and distribution checks have been undertaken. In order to limit the dataset size, the check-in and check-out dates have been restricted to the following time internal – January 2013 to January 2015. In order to gather a better understanding of how long customers usually stay in the booked hotels, following the distribution checks, a variable called ‘duration’, which is calculated as a difference between the check-out and check-in dates, has been created. Bookings with a duration greater than 14 days (99th percentile), have been excluded from the dataset to avoid issues pertaining to the skewing of results due to the impact of outliers. The final dataset consists of 22,105,405 rows and 25 columns.

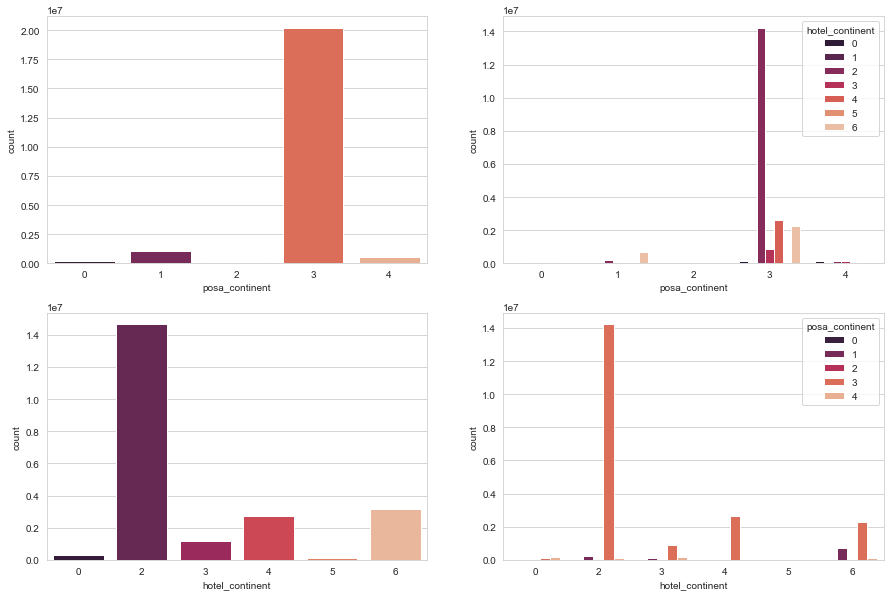


Figure 1: Number of Bookings by Continent

This chart above displays the frequency of bookings by traveller origin (“posa\_continent”) and traveller destination (“hotel\_continent”). Maximum bookings originate from continent 3 and are for hotels located in continent 2. The top right chart in the graph displays the breakup of hotel continents that travellers from an origin continent visit. As the charts in the left quadrants confirmed, travellers from continent 3 prefer to travel to continents 2,4,6, and 3 (domestically). The chart on the bottom right quadrant displays the breakup of traveller origin locations by destination continent. Since the majority of Expedia’s customers are located in continent 3, continents 2, 4, 6, and 3 have the highest number of bookings.

According to the box plot below, which plots the booking duration of a transaction against the destination continent, continent 4 has the highest median duration of stays and continent 2, the lowest.

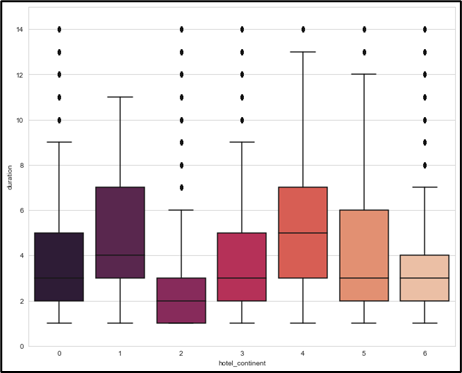


Figure 2: Duration of Bookings by Continent

The top figure below, displays the frequency of bookings made using mobile devices (left) and reservations made as part of a package (right). Approximately 20,000 bookings are made through mobiles and 25,000 as part of a package, which is roughly 14% of all bookings made during this period. Additionally, the chart on the bottom suggests that channel 9 is the most effective channel used while making a reservation.

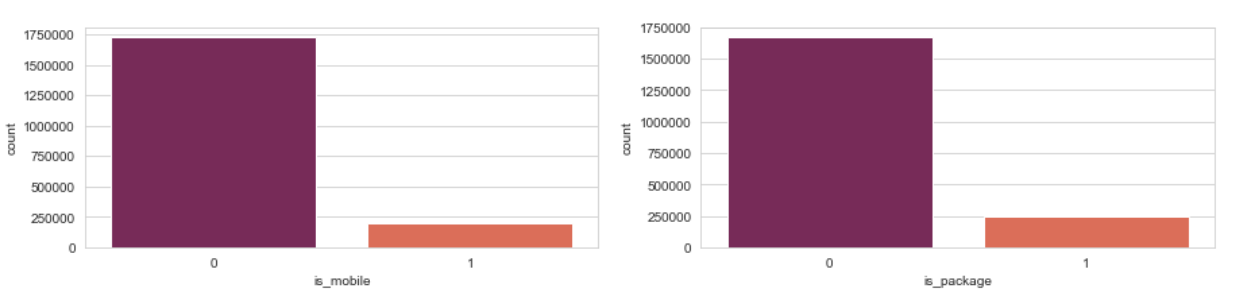


Figure 3: Mobile Bookings & Packages Booked

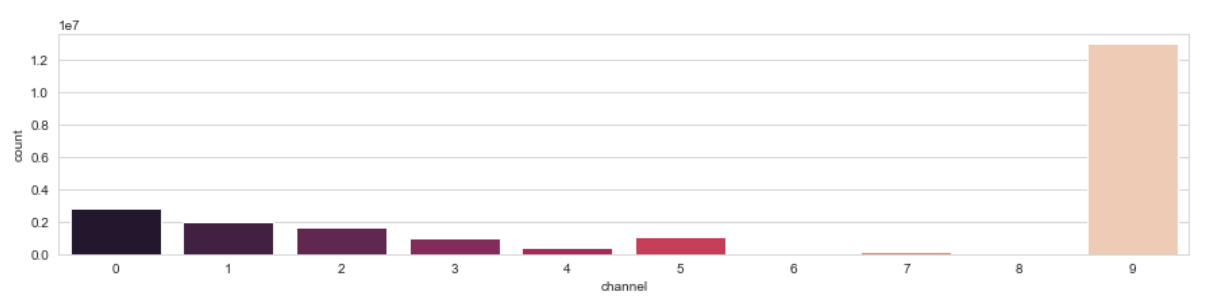
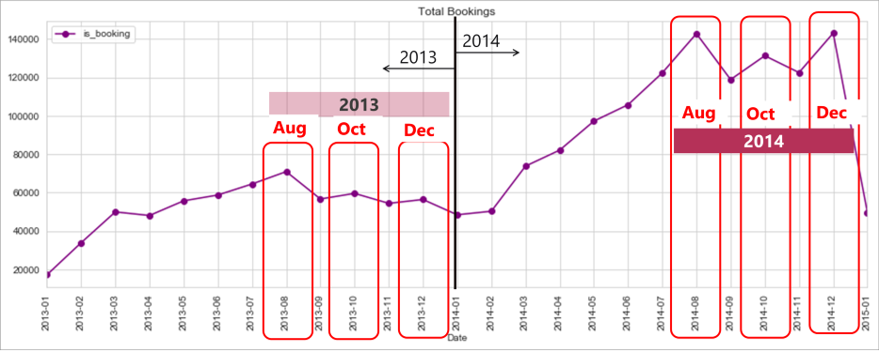


Figure 4: Booking Channel Used

The number of total bookings shows as increasing trend over time, with bookings peaking in the months of August, October and December. The number of bookings range from 18,000 (Jan 2013) to 150,000 (Dec 2014), with a median value of 60,000 bookings. This suggests that Expedia’s popularity has been increasing over the two-year time span.



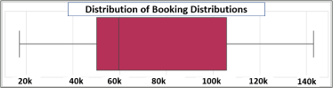


Figure 5: Distribution of Bookings by Time

The figure below displays the frequency of bookings for each hotel cluster, as grouped by Expedia. Clusters 91, 48,41, 65, and 42, appear to be the 5 most popular hotel clusters and cluster 74, the least popular.

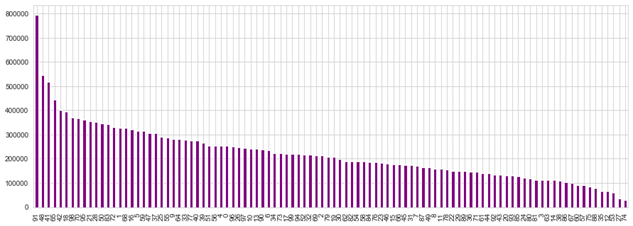


Figure 6: Number of Bookings by Hotel Cluster

The correlation matrix below displays the correlations between the top 8 highly correlated countries where hotels are booked. This chart could present valuable marketing strategies to Expedia, in terms of recommending holidays to customers.



Figure 7: Bookings Correlation between Countries

It appears that countries 8, and 47, in particular exhibit high correlations with countries 1,5, 50, 182, 185 and each other. Most other countries listed in the chart also display correlations with each other (>0.7), which can potentially lead to good cross selling strategies.

# Analytical Techniques

## Hotels Booking Prediction

By building machine learning model to predict whether customer behaviors would lead to a hotel booking, Expedia can better understand the key attributes that determine successful hotel booking. In addition, uncovering attributes that are important to customer behavior is critical to develop efficient marketing strategies. The following steps show how we accomplish the prediction and feature analysis:

1. Data Preparation
2. Feature engineering
3. Check for Multicollinearity
4. Training, testing dataset split and K-Fold cross validation
5. Building and training ensemble stacking model
6. Evaluation
7. Feature importance

***Step 1 Data Preparation***

Due to memory limitation exists in our personal machines, we are unable to train all the dataset for model. Hence, we filter out top 1000 active customer, i.e., those who made the highest number of bookings during the selected time interval. Eventually, there are 280,958 rows in the dataset. Next, we eliminate some columns that are redundant for the model. For instance, we remove all continent, country, city columns, since the values in these features are encoded so that we won’t know the real name of city or country even after analysis. Besides, there is another alternative feature, distance between user and hotel, to present geographical attribute much efficiently. We also create a new feature called “have kid”. Instead of using number of kids to predict booking of hotel, binary feature that divides all customers into “traveling with kid” and “traveling without kid” gives general view of how kids influence customer behavior of hotel booking. Afterwards, there are 12 columns remained in the dataset.

***Step2 Feature Engineering***

Apart from binary features, there are continuous variables and categorical variables in our dataset. All continuous variables have skewness and are not in the same range. Therefore, we transformed and scaled all of them. For categorical variables, because they are all encoded with numerical format, we just need to make sure they are presented in category type in the dataset, and they will be able to be trained by algorithm. Namely, make all categorical variables factorized.

***Step3 Check for Multicollinearity***

Since multicollinearity between any pair of continuous variables violates assumptions of most of algorithm, we need to assure and adjust them before going to model step. Fortunately, there is no high correlation (>0.7) between any pair of continuous variables as shown in below heatmap.



Figure 8: Heatmap of correlation between variables

***Step 4 Training, Testing Dataset Split and K-Fold Cross Validation***

We split training and testing dataset in accordance with the ratio of 7:3. As we only selected part of the original dataset, we leverage on K-Fold cross validation to examine how model performs generally on unseen data sample that not used during training process. We set K equals as 6, which is in the typical K range from 5 to 10 (Dekanovsky, 2020).

***Step 5 Building and Training Ensemble Stacking Model***

We set column “is\_booking” as target variable and other columns as features. We use tree-based algorithms to build an ensemble stacking model. The first level composed of Extra Trees, Random Forest, Gradient Boost, and Ada Boost learners. After training in the first level model, each learner will generate a score for a row, and we concatenate the scores to serve as second level training dataset. The second level is XG Boost, which is a renowned powerful learner on Kaggle and in the industry (Fan, 2020).

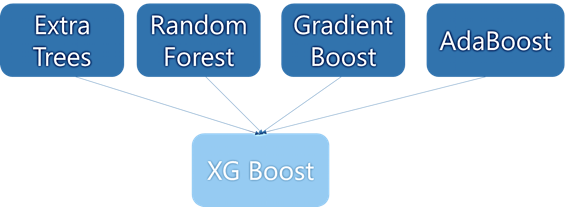


Figure 9: Model structure

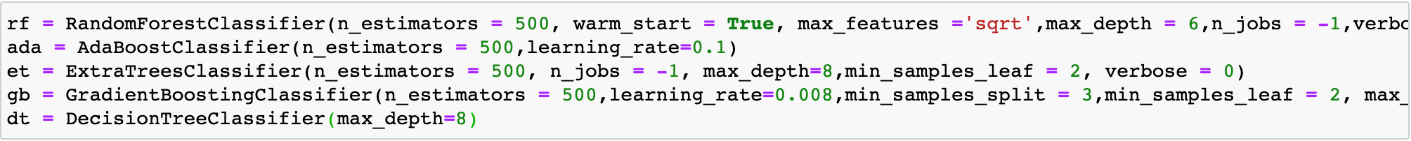


Figure 10: First level model

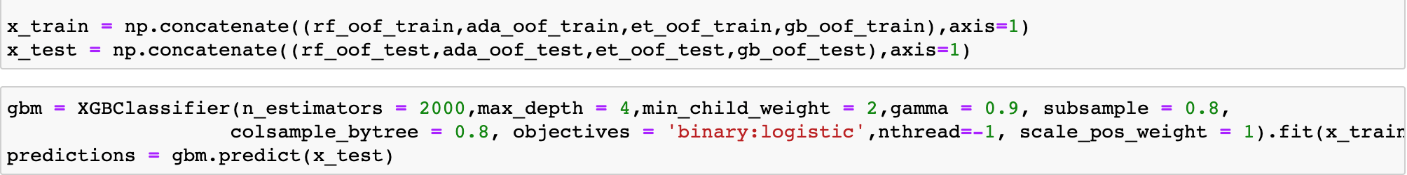


Figure 11: Concatenation of first level outcome and second level model

***Step 6 Model Evaluation***

After training and predicting the model using training dataset, we evaluate the model with reference to the accuracy rate. As shown in the picture below, the model accuracy rate is 94.11%, indicating good model performance.

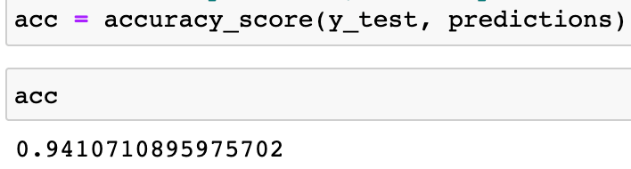


Figure 12: Model accuracy of 94%

***Step 7 Feature Importance***

In order to better understand which feature influences customer behavior the most, we used feature importance technique on XG Boost model we just trained. We found that distance between destination and user is the most influential factor. In the future, Expedia can go further with this factor to build managerial strategies. For instance, the company can propose long-distance and short-distance trip campaigns in different markets.

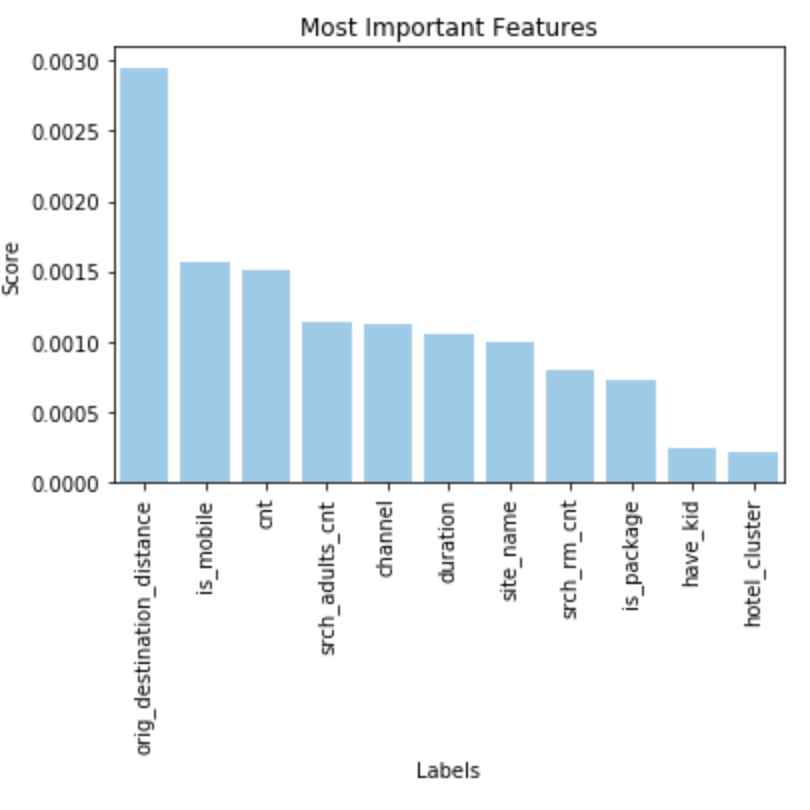


Figure 13: Bar graph of feature importance

## **Market Segmentation**

For the market segmentation, Customers’ behavioral data is used to divide them into different segments, based on their frequency, recency, and total number of rooms booked during the two-year period. The more defined subgroups would allow Expedia to reach the right people and provide more appropriate service for their customers.

***Step 1 Data Preparation***

Frequency, recency and monetary are three aspects that normally used for market segmentation, while the price of each booking event is not available in this dataset. Therefore, the *number of rooms* specified in the booking is used to indicate the value of monetary, based on the assumption that the more rooms an order had, the higher the booking price. Besides, we also try to apply *duration of the stay* and *duration\*total rooms* as indicators for segmentation*,* which are not qualified though. The frequency is the total bookings per customer during the two-year period and recency is referring to the time since last order of each customer.

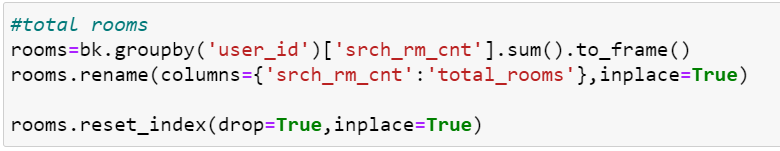


Figure 14 Calculation of Total Rooms

***Step 2 Feature Engineering***

Given that the K-Means clustering algorithm uses the Euclidean distance, three target variables 'total rooms', 'recency' and 'frequency' should not have severely skeweddistribution, so we perform log-transform and normalization for three variables, to ensure they become normally distributed. As a result, the skewness value in three features decreases down to around 1.

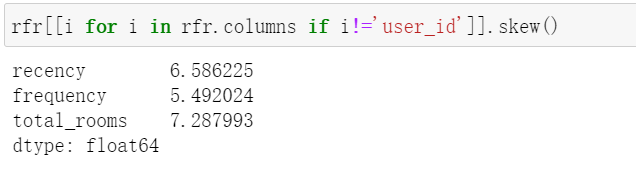


Figure 15 Original Skewness

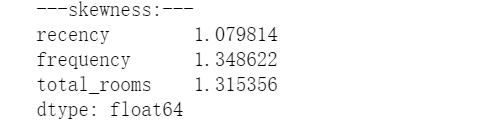


Figure 16 Skewness after Processing

***Step 3 K-Means Clustering***

K-Means clustering model is applied to implement market segmentation, and the final K-Value is selected based on the elbow method in *WCSS* score vs K-value plot (Mahendru, 2019), so as show in the chart below, we choose 4 as the k-value and assign customers into four groups. *WCSS* is the sum of squares of the distances of each data point in all clusters to their respective centroids (ODSC, 2018). The idea K-Means clustering is to minimize the sum.

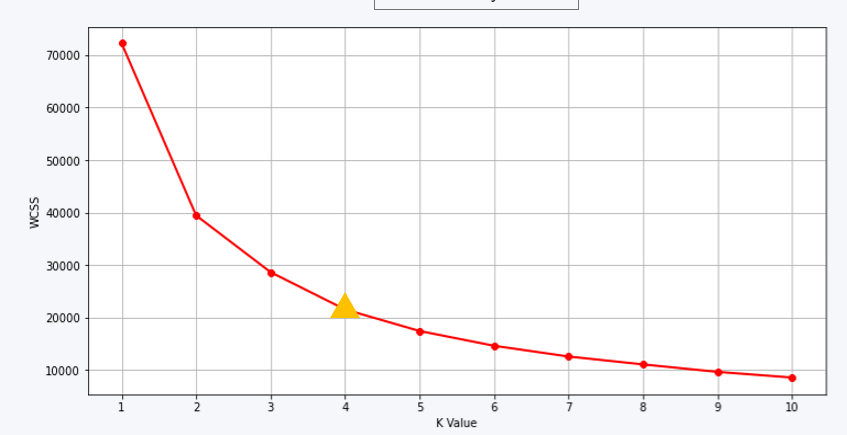


Figure 17: WCSS score vs K-value plot

These four groups, which are high-value customers, potential customers, new customers, and low-value customers, are named by their average values in three aspects.

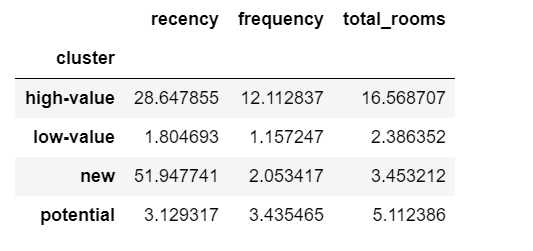


Figure 18:Avg. of Three Aspects of Four Clusters

***Step 4 Data Interpretation***

Among all customers have booked from Expedia, there are only 13% of the customers assigned into the high-value group, but about 33% of the customers are grouped to potential customers. At the same time, high-value customers contributed the highest number of bookings, up to 1,426,712, followed by potential customers, who had 925,181 bookings for the past two years.

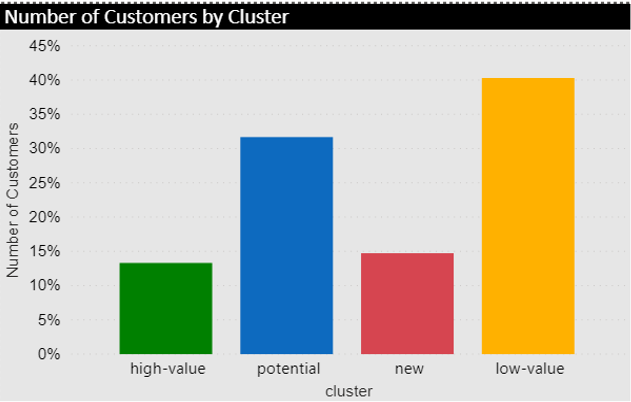


Figure 19:Number of Customers by Cluster

Besides, as can be seen from the bubble chart, where the value of total rooms is presented by the size of bubbles, high-value customers have the highest value in terms of total rooms and frequency, while the new customers have the highest value in recency, form which Expedia could make marketing campaigns specifically according to the characteristics of each segment.

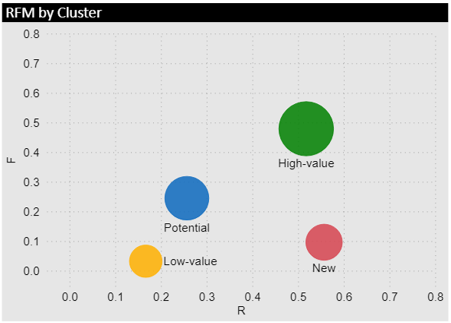


Figure 20:Characteristics of Four Clusters

## Market Basket Analysis

For the purpose to figure out which hotel clusters were booked together, we made a market basket analysis based on the four customer segmentations. We used SAS to make the analysis and the report settings are minimum confidence level at 10, support at 5%, and all results sorted by lift.

**Part 1. Results of Market Basket Analysis**

* High-value Customer

There are 1,426,712 records and 108,291 customers in the high-value customer group. The Average booking in this group is 13 that is the highest among the 4 segments. The result shows that there are 25 rules in the group, and the length of the relations includes 2,3, and 4. The central point in this group is 48,50 and 46.

Here is an example of rule interpretation, the first rule ‘50&32à39’ has 40.11% confidence, 1.49% support, and 4.49 lift. The result means 50,32, and 39 clusters occur together as 1.49% of all bookings for customers. The confidence level means once 50 and 32 clusters occur together then the probability of occurring cluster 39 is 40.11%. In the end, the lift is greater than 1, which means if a customer booked 50 and 32 before and the recommendation on cluster 39 would be effective.

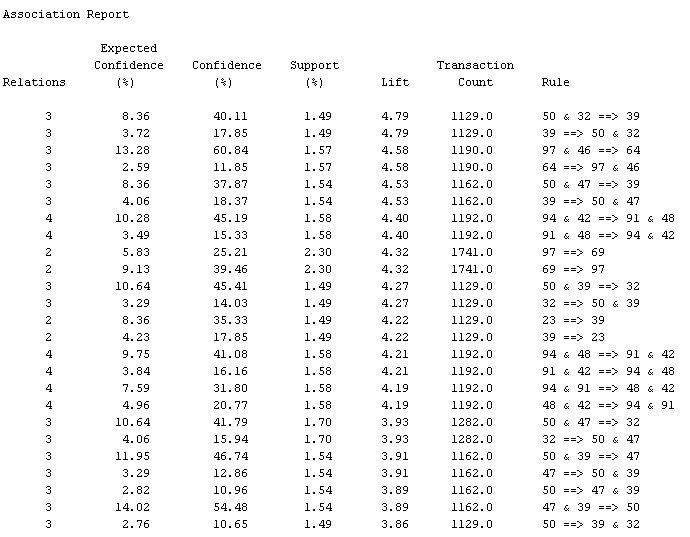


Figure 21: MBA for High-value customers

* Potential-value Customer

There are 925,181 records and 257,809 customers in the potential-value customer group. The average number of bookings made by each customer in this group is 4. The results show that there are 25 rules in the group, and all the length of the relations is 2. The central point in this group is hotel cluster 48 and 91.

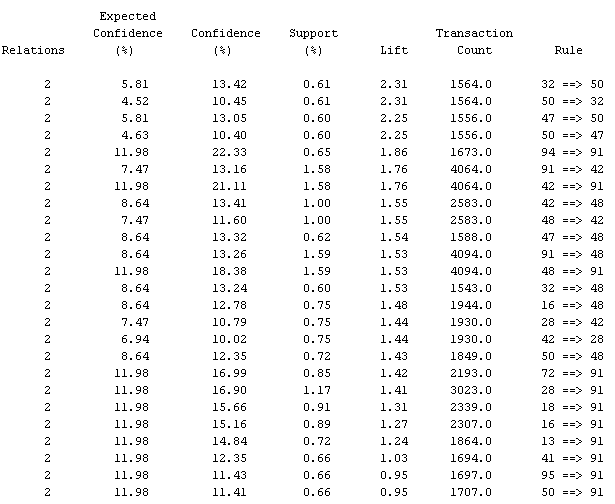


Figure 22: MBA for Potential Customers

* New Customer

There are 259,666 records and 119,872 customers in the new-value customer group. The Average booking in this group is 2. The result shows that there are only 3 rules in the group, and all the length of the relations is 2. The central point in this group 91.

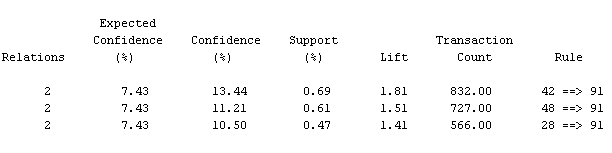


Figure 23: MBA for Low-value customers

* Low-value Customer

In addition, there are no relevant association rules in low-value group because each customer in this group only booked once.

**Part 2. Result Comparison**

From the comparison, the result of MBA analysis for High-value customers appears to be better than that of the other 2 clusters because there are more association rules returned and more booking records for each customer. It is worth noting that the lift value in some rules is smaller than 1 and those rules is considered as not effective for recommendation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **AVG Confidence** | **AVG Support** | **AVG Lift** |
| High Value | 28.15 | 1.73 | 3.33 |
| Potential | 13.78 | 0.87 | 1.53 |
| New | 11.71 | 0.59 | 1.57 |

Table 1: Comparison Table for Market Basket Analysis Results

## Recommendation System

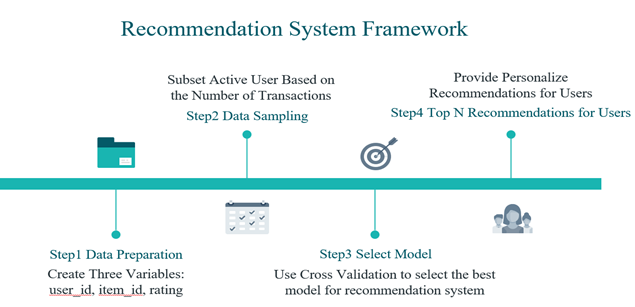


Figure 24: Recommendation System Framework

The above flow chart shows that there are four steps to build the recommendation system:

1. Data Preparation
2. Data Sampling
3. Select Model
4. Top 3 Recommendations for Users

***Step1 Data Preparation***

In order to use *surprise* package, which is a python packages for recommendation system, it requires three variables: user ID, item ID, and rating. For user ID, our group use user\_id in the Expedia Hotel dataset. Regarding item id, we select hotel\_cluster as our item id. However, because the dataset does not include rating information, our group create a derived variable by using the number of transactions of each user.

We group by user\_id and item\_id and the result is the following photo. The dataframe has 9,959,198 rows and 3 columns.

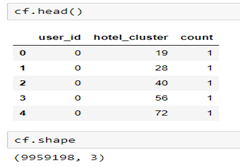


Figure 25 Recommendation System Dataset

***Step 2 Data Sampling***

There are two reasons why our group implement data sampling. Firstly, in order to build an accurate recommendation system, we select the active users because these users have higher transaction frequency, and they make the model more accurate. The other reason is that due to the computation limitation of our laptop, it is easier to build the recommendation with a small dataset. Hence, our group filter out users with transaction record range between 3 to 10 (do not consider outliner) and then randomly select 10,000 users for building the model. The dataset where we select 10,000 from is the below table.

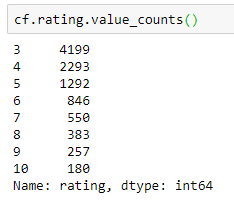


Figure 26 Rating Distribution

***Step 3 Model Selection***

There are many recommendation models. For our project, we would like to focus on K-Nearest-Neighbors (KNN) instead of SVM because KNN is used as multi-class classifier and SVM is often used as binary classifier based on research papers (Grcar, 2006) .Our group implement four different KNN algorithms: KNNBaseline(), KNNBasic(), KNNWithMeans(), KNNWithZScore(). We measure the accuracy with RMSE and three times cross validation. The result shows that KNNBasic() has the smallest RMSE like the table below.

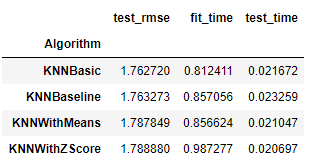


Figure 27Model Evaluation Results

***Step 4 Top 3 Recommendations for Users***

Our recommendation system will provide users three hotel cluster based on the dataset. For example, if user, 924020, visits our platform, then we will recommend the user hotel cluster 55 (the forecasted rating for this hotel cluster is 10.0), hotel cluster 9 (the forecasted rating for this hotel cluster is 7.0) and hotel cluster 58 (the forecasted rating for this hotel cluster is 6.0) like the below photo. We believe that this recommendation system could drive the revenue of Expedia Hotel since it does not have this service currently.

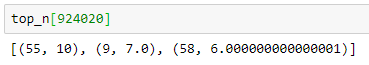


Figure 28 Recommendation System Demo Example

# Conclusion

In this report, we took a customer-centric perspective to leverage on customer analytics techniques in order to understand further Expedia’s customer behavior with the intention to extrapolate the analysis to further enable Expedia with more customer-focused service offering.

With reference to customer lifecycle as denoted by (Linoff & Berry, 2000), to help Expedia better facilitate customer acquisition, we deployed machine learning models and arrived at a 94% accuracy rate in predicting whether customers are likely to make a booking based on available features. With the significance for businesses like Expedia to acknowledging the differences within their customer base, we leveraged on RFM technique to segment the customers into four unique segments to better tailor promotional efforts with more suitable marketing strategies.

With a plethora of hotels listed in Expedia’s network, our analysis from market basket analysis (for high-value customers in particular) offers insightful findings for them in the hidden association among seemingly-stand-alone hotels which could be further utilized for marketing purposes. Furthermore, analysis on recommendation system based on KNN model illustrates the possibility of further developing existing customers to effectively offer contents that more accurately fit their needs.

Subsequent analysis on this topic should also focus on decoding the anonymized locational data with the assistance of external data to generate more sensible insights with hotel and customer locations. Customer ratings on hotels listed in Expedia could be appended to the analysis to provide more context in customer feedback on different hotels to more efficiently derive customized offering. In hotel booking area especially when it is booked for leisure travel purposes, customers generally make few bookings every year. This point is commonly observed in our analysis as with only two years of data. Future analysis could focus on a specific active customer segment or include more years of data to support more relevant analysis.

# Bibliography

Dabbura, I. (n.d.). *k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks*. Retrieved from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a.

Dekanovsky, V. (2020). *Complete guide to Python’s cross-validation with examples*. Retrieved from towardsdatascience: https://towardsdatascience.com/complete-guide-to-pythons-cross-validation-with-examples-a9676b5cac12

Fan, P. (2020). *XGBoost, LightGBM, and Other Kaggle Competition Favorites*. Retrieved from towards data science: https://towardsdatascience.com/xgboost-lightgbm-and-other-kaggle-competition-favorites-6212e8b0e835

Grcar, M. (2006). *kNN Versus SVM in the Collaborative Filtering Framework.*

Heng, K. (2019, Oct). *Predicting Hotel Bookings on Expedia*. Retrieved from towards data science: https://towardsdatascience.com/predicting-hotel-bookings-on-expedia-d93b0c7e1411

Linoff, G., & Berry, M. J. (2000). *Mastering Data Mining: The Art and Science of Customer Relationship Management.*

Mahendru, K. (2019). *How to Determine the Optimal K for K-Means?* Retrieved from Medium.com: https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb

ODSC. (2018). *Unsupervised Learning: Evaluating Clusters*. Retrieved from ODSC - Open Data Science: https://medium.com/@ODSC/unsupervised-learning-evaluating-clusters-bd47eed175ce

Raeder, T., Raeder, T., Chawla, N. V., & Chawla, N. V. (2011-04). Market basket analysis with networks. Vienna: Springer Vienna.

Wikipedia. (2020, October). *Expedia*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Expedia

# Appendix 1 Details on Expedia hotel booking dataset

* *Location and destination*

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Feature Description** | **Data Type** |
| user\_location\_country | ID of the country the customer is located | int |
| user\_location\_region | ID of the region the customer is located | int |
| user\_location\_city | ID of the city the customer is located | int |
| destination\_id | ID of the destination where the hotel search was performed | int |
| destination\_type\_id | Type of destination | int |
| hotel\_continent | Hotel continent | int |
| hotel\_country | Hotel country | int |
| hotel\_market | Hotel market | int |
| orig\_destination\_distance | Physical distance between a hotel and a customer at the time of search. | float |

* *Hotel search settings by customers*

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Feature Description** | **Data Type** |
| srch\_ci | Check-in date | object |
| srch\_co | Check-out date | object |
| srch\_adults\_cnt | The number of adults specified in the hotel room | int |
| srch\_children\_cnt | The number of (extra occupancy) children specified in the hotel room | int |
| srch\_rm\_cnt | The number of hotel rooms specified in the searchsc | int |

* *Other features collected by the platform*

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Feature Description** | **Data Type** |
| date\_time | Timestamp | object |
| site\_name | ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk) | int |
| posa\_continent | ID of continent associated with site\_name | int |
| is\_mobile: | 1 when a user connected from a mobile device, 0 otherwise | int |
| is\_package | 1 if the click/booking was generated as a part of a package,0 otherwise | int |
| channel | ID of a marketing channel | int |
| is\_booking | 1 if a booking, 0 if a click | int |
| cnt | Number of similar events in the context of the same user session | int |
| hotel\_cluster | ID of a hotel cluster | int |