Five College DataFest 2017
April 2nd, 2017
Team: Normally Distributed

MARKETING ANALYSIS FOR EXPEDIA

We focused on marketing for Expedia to be able to better target as well as customize channel communications to different customer types. First, we analyzed the global customer distribution of Expedia. In order to this, we group by each country and take the log10 of sum of user activity. From this graph, we can infer that the area between light blue and dark blue may have a lot of potential for future growth. They have a relatively lower user activity than North America but relatively higher than the rest of the world.

In order to better target different types of customers, we categorize 6 types of customers to see their probability of booking each price range of hotels. These 6 types of customers are: People who are traveling less than 800 miles, between 800 miles to 2000 miles,above 2000 miles with/without kids. There is a clear upward trend for the high, VH price hotels when people travel far away. Each even number clients are clients who bring kids and we can see that those people have a higher probability booking higher price hotels. These results provide Expedia more information about how to recommend more suitable price range hotels for customers.

The tree on the right shows the average number of days people stayed based on certain characteristics we can see from the data. In our model, we did not include variables that asked about geographical location because the data set only mentioned the country the hotel is located in. The root MSE was 2.5 days.

We also attempted a cluster analysis to provide further context to our understanding of Expedia's customers. Our silhouette plot shows the average dissimilarity when performed for 2 to 10 clusters. Because this was run on the mini_clicks data table, there was weak structure found, even for the best number of clusters (clusters = 3, sil. width = 0.17). However, there are some interesting differences between clusters on this sample. The interactive graph represents the relationship between our variables in a lower dimensional space with t-distributed stochastic neighborhood embedding (t-SNE) for dimension reduction. Some interesting features include: Cluster 1 (blue) was mostly users from the United States, flying domestically and looking for unbranded hotels; cluster 2 (yellow) is mostly Canadians planning trips 2+ months out and aiming for popular; cluster 3 (grey) is mostly US fliers going shorter distances and planning 3-6 weeks in advance. These clusters might have a stronger relationship if run on a larger sample.

Next, we look at what drives the purchase rate of a destination. We use a random forest model to predict the conversion rate of a destination based on its popularity scores of different areas. We find that being popular for business trips, sport activities, beaches, and historical sites are the most influential factors for hotel purchase rate of a destination. The cross-validation MSE is 0.018 and the range of the conversion rate is from 0 to 0.25. We also examine Yelp data to find common patterns in both Yelp's and Expedia's customers. We find that places with more businesses on Yelp tend to be rated higher in popularity for business trips, family, food, and social bars. However, the correlation assumptions are not met so we should infer with caution. The results help determine the types of destination to advertise to customers.