A Business Opportunity:
Targeting Expedia's Niche Market in Travel Packages
Via Analytical and Predictive Modeling

Subashini Sridhar, Ji Won Chung, Ji Young Yun, Zainab Rizvi

Abstract

The dataset for the ASA 2017 Datafest competition was provided by Expedia Inc., a travel company that primarily runs travel fare aggregator websites. The dataset includes over 10 million user records of searches and purchases through various Expedia websites. This paper conducts a machine learning analysis via a classification decision tree to identify potential customers who do not purchase a travel package but are similar to those who do. The paper then narrows down on the countries a group of potential customers is most likely to travel to as well as the types of hotels. The paper investigates the group's preferences by looking at factors such as hotels' price range, star rating, and brand. These features create a predictive model that help suggest to Expedia the niche markets to focus on to convert non-consumers to those who purchase travelling packages and as a corollary to increase revenue.

I. Background and Significance

The dataset for the ASA 2017 Datafest competition was provided by Expedia Inc., a travel company that primarily runs travel fare aggregator websites. The dataset includes over 10 million user records of searches and purchases through various Expedia websites. We are interested in understanding the types of customers that purchase vacation packages through Expedia, providing actionable recommendations to Expedia in increasing the number of packages booked through their website, and converting customers who book through the website into package buyers.

II. Data and Methods

The explanatory variables used for analysis include duration, daysAhead, orginDestDist, and party_total. The daysAhead variable represents the number of days ahead of travel the customer is booking; it was created by calculating the difference between check-in date specified in the customer search and the current date of booking. The physical distance in miles between a hotel and a customer at the time of search is represented by originDestDist. Party_total is a numerical variable representing the total number of people on the package. It was calculated by adding the number of adults specified to occupy the hotel room and the number of (optional) children specified to occupy the hotel room. The dependent variable is is_package, a binary categorical variable that indicates if a customer booked a package or not via Expedia. A package purchase is defined as a hotel booking combined with a flight ticket and/or a car rental reservation.

The research takes a subset of the data set in order to identify the potential group of customers Expedia can market. This set is a binary categorical variable called is_booking which represents the Expedia customers who were involved in any type of booking. The data is further parsed to exclude customers travelling less than one day, those not making bookings at least one day ahead travel, and those with data unavailable for origin to destination distance. In total, 680,615 Expedia customers are used to fit and analyze the model.

A classification decision tree is used to model is_booking as a function of daysAhead, orginDestDist, and party_total. This tree model includes information regarding customer preferences and background that is easily interpretable for representatives at Expedia. To measure the performance of our model, we predict the model onto our dataset and draw a confusion matrix to assess the accuracy rate.

III. Results and Discussion

The classification decision tree model highlights that only 3% that of all the customers in our dataset that book through Expedia buy a package and 97% of them do not make package purchase. According to the model, the primary deciding factor for an Expedia customer in booking a package is the duration of the travel. If the duration of travel is less than 2.5 days, our model predicts it is highly unlikely a package will be bought 97% of the time.

For customers with a travel duration greater than 2.5 days, the tree splits further based customer's preferences in duration of travel, distance from origin to destination, the number of days ahead of travel booking is made. Two unique group of customers were identified by traversing the decision tree on customers with a duration greater than 2.5 days. The first group, identified as Group A, is typically composed of party sizes greater than or equal to six people, make reservations at least 322 days ahead of time, travel for less than five and half days, and travel for a distance of greater than or equal to 2,142 miles. The probability

for an Expedia customer that matches the conditions in Group A purchasing a package is 54%.

Customers in Group B share all the same characteristics as those in Group A, except their distance of travel is less than 2,142 miles; and this becomes the turning factor for the group's decision in typically not buying a package. Refer to Figure 1, a subset of nodes from the decision tree noted in Appendix A, that differentiates Group A from Group B.

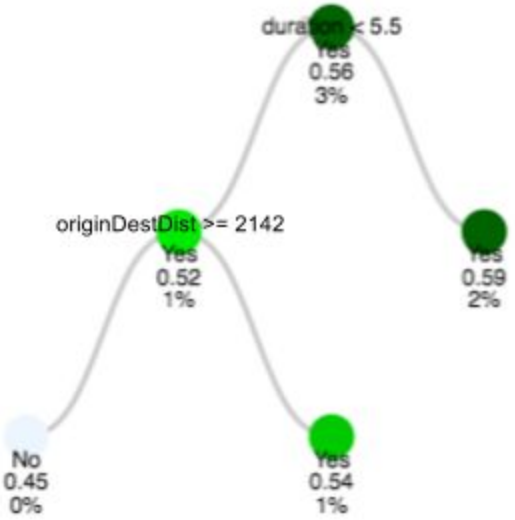


Figure 1: Subset of Decision Tree that Represents Potential Package Buyers

Predicting the classification decision tree model onto the dataset tells us that the tree correctly classifies customer's Expedia booking 91.8% of the time. Refer to Appendix B for the confusion matrix and accuracy calculations. At first glance, it may seem the model is performing exceptionally well, but in truth it is a little misleading because we trained and tested the model on the same set of 680,615 observations. In other words, 100 – 91.8 = 8.2% is the training error. The training error is often more optimistic and tends to underestimate the test error rate. Regardless, we can still draw insights for Expedia from customers in Group A and Group B.

The set of customers that are most prime for conversion, or to buy a package, is Group B because their traits are similar to Group A. This paper further examines the differences between Group A and B to find the factors that affect a customer's decision to buy a package. The analysis focuses on the countries that Group B is most likely to travel to as well as the types of hotels, looking at factors such as price range, star rating, brand, it prefers. This insight will allow Expedia to customize packages for Group B and capture this critical market segment.

We divided the travellers from groups A and B into a total of four clusters based on two factors: those who traveled from the United States (U.S.) to foreign countries and vice versa. We focused on the U.S. because U.S. travellers are Expedia's largest consumer market and what the data is mainly based on. We then ranked each cluster based on which country was the most popular travel destination or origin of travel. Refer to Appendix C for a visualization that summarizes this data.

By basic filtering of the dataset by travel destination, we found that Expedia had a solid market in the countries that people traveled to and from the most. The top countries that had the most number of people traveling to the United States (U.S.) were Canada, Germany, and Mexico. The top three countries people chose to travel from the U.S. are Mexico, Dominican Republic and Puerto Rico. This indicated that Expedia did not need to focus on expanding their traveling packages in these countries because they already have a profitable revenue and solid consumer base.

To identify the countries that have potential consumers for travel packages, we focused on identifying countries people do not often travel to and do not buy packages in. The Dominican Republic and the United Arab Emirates were two of these countries.

More focus on marketing travel packages to these countries would increase revenue.

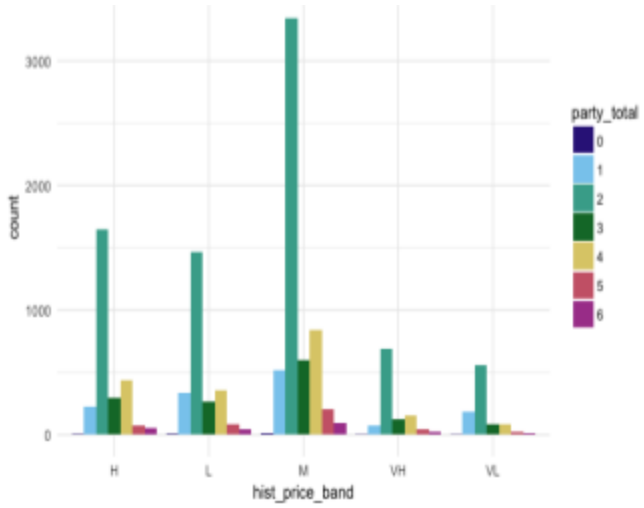


Figure 2: Bar Chart Comparing Customers who Purchased Packages

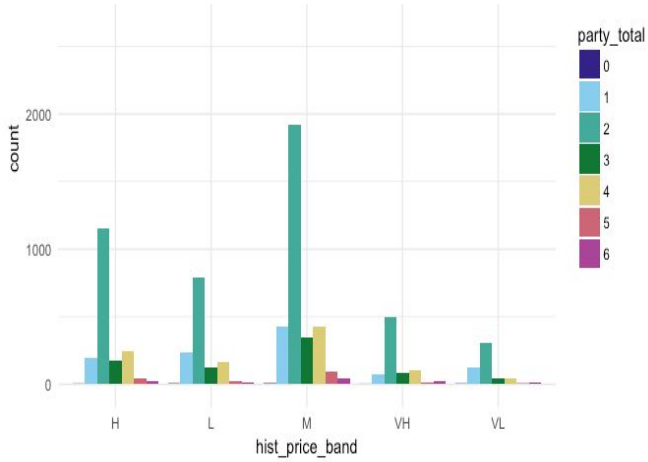


Figure 3: Bar chart Comparing Customers who did not Purchase Packages

The bar charts from Figure 2 and 3 depict the number of hotels with various price ranges booked by parties of different sizes. The x-axis represents the price range of the hotel: High, Low, Medium, Very High, Very Low. The y-axis represents the number of hotels booked by customers. The hotels with medium price range are the most popular. People tend to choose branded hotels with star ratings of 4 or 5. In addition, groups with smaller number of people are likely to choose hotels with lower range prices. Comparing the two charts shows that there are more groups of 2 in the group of people who bought packages. This information is important because understanding consumer behavior enables Expedia to cooperate with potential hotels to maximize profits, better serve customers' needs and wants, and develop appropriate business strategies to increase the number of people who buy packages.

IV. Conclusion

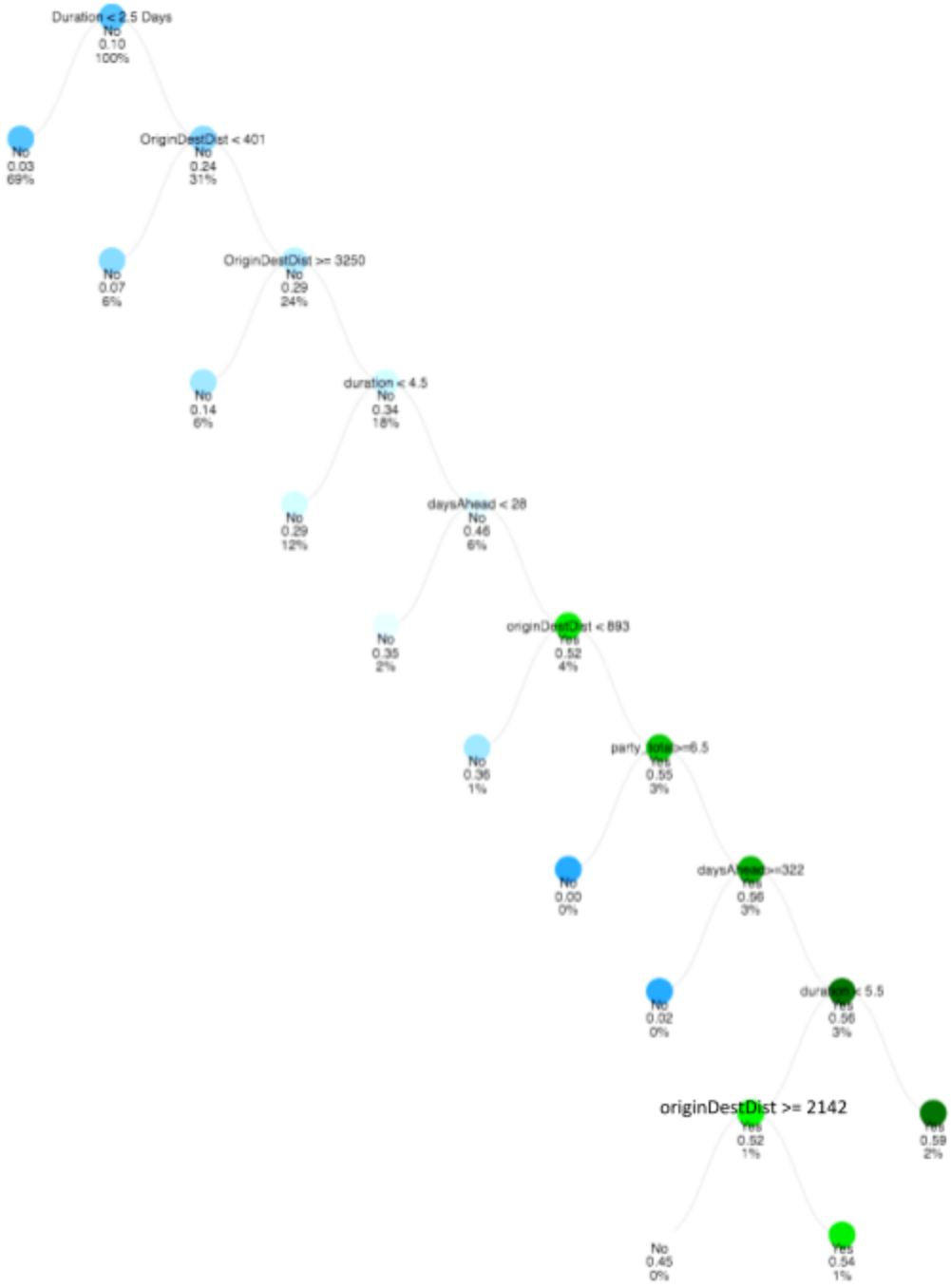
For Expedia to capture a profitable market, the company should do the following: (1) tier hotel packages according to party sizes and (2) focus on different regions for US and non-US travelers. The following was a real sample case that shows a concrete estimate of potential revenue Expedia could earn. An Expedia search on 04/02/2017 that mimicked the behaviors of Group B was conducted with the following criteria: (1) travel from New York to Cancun (2) travel duration of six days (3) book 29 days in advance (3) travel from 05/02/2017 - 05/08/2017 (4) travel to a four star hotel. The discounted package including cars, hotels, and flights costed $1,111. The last node from Figure 2 demonstrated 20,785 people who were prime for conversion. Targeting this market segment would lead to an approximate 20,785 * $1,111 = $23 million revenue for Expedia.

References

"Expedia Travel: Vacations, Cheap Flights, Airline Tickets & Airfares." *Www.expedia.com*. N.p., 2017. Web. 2 Apr. 2017.

Schumacher, Aaron. "Interactive D3 View of Sklearn Decision Tree." *Popular Blocks*. N.p., n.d. Web. 2 Apr. 2017.

Appendix

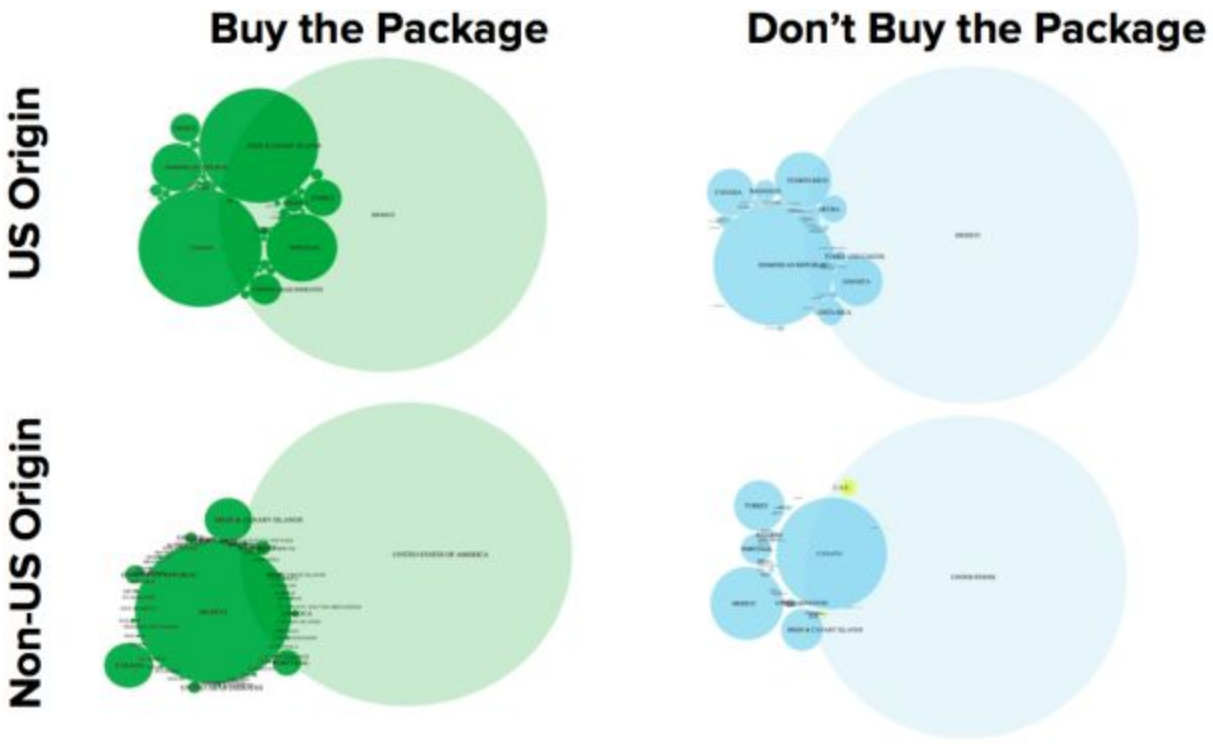


Appendix A: is the full model of the classification decision tree we created for our analysis.

pred_tree	Predicted:0 (package not bought)	Predicted:1 (package bought)
Actual: 0 (package not bought)	614674	57256
Actual: 1 (package bought)	7643	10402

((614674 + 10402)/(614674 + 57256 + 7643 + 1042)) * 100 = 91.8% Accuracy

Appendix B: Confusion matrix created to measure performance of the binary classification classification model. Model predicted on the entire dataset.



Appendix C: Cluster of countries depending on origin and decision regarding package