Theoretically Statistics

Azka Javaid, Andrew Kim, Brendan Seto, Jason Seto

For our analysis, our main concern was understanding how groups of users behave while using Expedia.com. From our one million observation subset of the data, we categorized all users into 4 categories: families (groups including one or more child), groups (groups including three or more adults), couples (groups of two adults without children), and solo (searches involving only one adult).

Analyzing Travel Flows of Users

We first wanted to understand the travel patterns of users from these four groups. We separated trips into domestic and foreign trips, with domestic trips being ones where the user country and the hotel country are the same, and foreign trips are ones where the user country and hotel country are different. We only wanted to look into foreign flights for all four groups and plotted the data in order to visually see the popularity of each type of trip (defined as a unique combination of hotel and user country).

User Data Aggregation

We observed that certain users viewed multiple hotels on the same check in date and hotel. Since this is likely to be the same person, we aggregated the dataset grouped by user id, check in date and hotel destination country. This dataset was used for all further algorithmic and predictive analysis.

Analyzing Groups of Travelers

We also wanted to see what times of the year are most popular for hotel bookings for each group of people, as well as how far in advance different groups of people search for hotels. From our analyses, it appears that family trips have a very clear seasonal trend, as activity is most prevalent in mid-spring, summer and winter, consistent with school vacations. Solo trips were the most uniform. We suspected that solo travelers could be broken into business and leisure travelers. To investigate this, we used UK Customs data for 2015 to analyze how purpose of travel changes over time. Business travel was found to be uniform across the year, while leisure travel peaked at around late summer. This suggesting that the travel purposes of solo travelers may reflect similar trends captured in the UK Customs data.

Also, we found that users of all four groups tend to search for hotels closer to their intended check in times. However, we noticed that solo travelers are much more likely to search for hotels closer to their check in times than travelers from the three other groups. We conducted an ANOVA test on the logarithmically transformed values of “advance” (difference in days between timestamp and check in time) to determine whether a significant difference between groups for the “average” variable exists. We found that a statistically significant difference did exist, and that the only statistically significant difference exists between solo travelers and the other three groups of travelers.

Predicting User Class

We modeled the class groupings against user and hotel attributes to find features that best predict the class groupings. We included explanatory predictors like mobile_use, package_use, channel, stay_duration, total_times (total time spent by the user on expedia), times (number of times user searched for that trip), advance (difference between search time and checking in time), ave_star (average star rating of all search queries), range_star (range of star ratings of all queries), ave_price (average price

rating of all search queries), range_price (range of prices of all queries), ave_dist (average distance between hotels in the area)
Range_dist (range of distance between hotels in the area), num_branded (number of brand name hotels searched for a trip) and num_bookings (number of bookings made by a user for that trip).

We partitioned data in 80/20 training and test set split on 500,000 observations. We modeled a random forest with 100 tree iterations (n = 100) and sampled 3 predictors for splitting at each node (mtry = 3). Our model accuracy was 59% with advance, total_times stay_duration, channel and ave_dist being the most important predictors.

Additionally, we used multinomial logistic modeling to observe the direction of the coefficient effects on predicting the user class. The table below summarizes key coefficient relationships in predicting user class.

Coefficients of Multinomial Logistic Regression (Relative Odds to Couple)

	totaltimes	stayduration	ave_dist	ace_price
family	0.0177	0.0343	0.04050	0.0729
group	0.0111	0.0359	-0.00863	-0.0000
solo	0.0175	0.0357	-0.02850	-0.1070

Predicting Hotel Class

We were then interested in clustering users in particular hotel clusters. We used the ave_price (average price of all hotels searched) and ave_pop (average popularity of all hotels searched) to create 4 hotel clusterings that grouped the hotels based on price and popularity. We then modeled this hotel_type predictor by user attributes which included predictors like use_mobile, use_package, channel, stay_duration, class, times, total_times, advance and is_domestic. We used a 80/20 model split on 500,000 observations. A random forest with 100 tree iterations and random sampling of 3 attributes on each node split was used for modeling. We observed advance, total_times, stay_duration and channel to be the most important predictors in predicting the hotel class with a model accuracy of 68%.

Additionally we used logistic modeling to observe the direction of the coefficient effects in predicting the hotel class. The table below summarizes key insights:

```
Call:
multinom(formula = hotel_type ~ use_mobile + use_package + channel +
    stay_duration + class + times + totaltimes + advance + is_domestic,
    data = train)

Coefficients:
           (Intercept) use_mobile1 use_package1  channel262   channel293  channel324 channel355  channel386
cheap_unpop   -2.883764  0.03626490   -0.7831582 -0.06267556  0.008591956 -0.06008198 -0.3674442 -0.06317201
exp_pop        1.205864 -0.09600456    0.2656267 -0.02337113 -0.083591991  0.17939635 -0.1248588  0.01074974
exp_unpop     -2.457691 -0.09226762   -0.3268995 -0.06198340 -0.169101043  0.07779995 -0.1999840 -0.19717635
           channel417 channel448 channel479  channel510 channel541 stay_duration classfamily  classgroup
cheap_unpop -0.4571720 -0.3287107 -0.1068308 -0.20259750 0.01542994   0.014831848 -0.08854804 -0.10985010
exp_pop     -0.3157622  0.1157097 -0.1337016  0.01908599 0.10840523  -0.015381291  0.10012890 -0.03521363
exp_unpop   -5.1782760 -0.1128841 -0.4298318 -0.22449666 0.04310945   0.007090674  0.12806650 -0.01320706
            classsolo       times    totaltimes      advance is_domestic1
cheap_unpop  0.1568080  0.01706796 -0.0006732798 1.548563e-05   0.05316056
exp_pop     -0.2346183 -0.30048366  0.0088726876 7.948709e-04  -0.07800061
exp_unpop   -0.2335024 -0.41322815  0.0125534709 8.271468e-04   0.64306472

Residual Deviance: 609530.2
AIC: 609656.2
```

Conclusions

We performed a random forest for our predictive analysis to predict the user class based on user and hotel attributes and then predicted the hotel class based on user attributes. We additionally used a multinomial logistic regression to observe the directional effects of the coefficients on the specified response predictor. Grouping by class allows us to understand certain individual behaviors, giving potential insights into advertising strategies. Grouping by hotel type allows us to see what kinds of hotels people click on given the users' characteristics. This will allow us to refine and develop personalized search algorithms.