General Presentation Script

—Intro slide—

Cristal:

Our final analysis focused on the importance of two variables, pickiness and vacationiness and their effect on the expedia is_booking variable. Where ”pickiness” is defined as a higher average number of clicks before making a decision to either abandon the hotel search or make a booking on Expedia. And ”vacationiness” is defined as the general likelihood that a destination is a vacation spot rather than just a business or other non-vacation spot.

—1st slide—

Cristal:

Our initial hypothesis was that airbnb’s market presence in a city was a notable source of variance on pickiness of a user. We chose three cities that exemplified the three most characteristic differences in market interactions between Airbnb and Expedia. These being: New York, Los Angeles, and Austin. These characteristic differences were that Airbnb was more expensive than Expedia in Austin, less expensive in New York and equally expensive in Los Angeles.

After our analysis of the mean scores of pickiness, we could confirm our hypothesis as we see in the three graphs for each city.

Younjoo:

However, these scores and cities also followed a general trend of large cities or vacation-worthy spots, so we had to test a city that was not a large city nor a vacation destination that had a similar airbnb vs expedia profile to New York and closer population to Austin; we found this city to be Seattle. Exploratory analysis on Seattle showed us that there are significantly larger sources of variance on is_booking than airbnb.

Our new hypothesis is that vacationiness is that larger source of variance. However, in our analysis of airbnb data we found that airbnb rentals are a source of variance in the hotel marketplace, which suggests that if we had a larger data set we could make significant correlations to users Expedia bookings and research further.

—2nd Slide—

Zack:

In our exploratory data analysis we noticed that the is_booking variable by itself is misleading because if someone searches for a hotel for three days without booking and then ends up booking on the third day, this will count as one booking and two not-bookings. The cnt variable only addresses this sort of issue within internet sessions. To address this issue we aggregated the data based on user-id and check-in date of hotel

searches. In the aggregation we include max(is_booking) and count(is_booking) where count(is_booking) represents pickiness score and max(is_booking) tells us whether the user made a booking or not across the aggregated data. max(is_booking) is our response variable. Our assumption based on the large discrepancies in pickiness of users who book vs don't book in our exploratory data analysis was that pickiness would be a large source of variance on the response variable. We then populated data with all of the viable predicting variables from the user data set, but aggregated with new aggregated variables included when necessary or probable and we ran this data in a random forest hoping to train a model that predicts is_booking better than 85% where 85% was an arbitrary number that we hoped to surpass. Without any coercion we trained a model scant of 80% predictive power. The random forest told us that the largest source of variance and predictive power is the pickiness variable count(is_booking) followed by an aggregated sum(cnt) variable. To our surprise only a few variables from the dest dataset were even negligibly useable as predictors of our response variable. With respect to prediction variables indicative of vacationiness, much better were the user_data variables for size of party, rooms sought, and star_rating of hotels searched. This leads us to believe that we could build a better random forest with more aggressive training on the full data sets to better predict which destinations will lead to larger party sizes, or even the inclusion of such an average party size in the destinations data file. If the dest variables held more predictive power, we would feel confident in saying that our training model suggests that vacationiness is as important as pickiness in predicting is_booking or even pickiness itself. However, our model did pick up on party_size and rooms sought as a high source of variance and a good predictor, so we still believe that vacationiness is important, but

more research should be done to predict a destinations vacationiness instead of users search behavior predicting the destinations vacationiness. The few dest variables that held negligible predictive power were related to luxury, food, beer, and natural features.

To sum it all up, pickiness is a very important predictor of is_booking and vacationiness appears to be a good predictor of pickiness and thus by the transitive property should also be a good predictor of is_booking too!