Datafest Reflections

What we did, did well, and not so well:

We initially wanted to test for a correlation between the expedia data and publicly available airbnb data. Because we weren't given price information about the hotels listed in expedia, we instead wanted to see if the discrepancies in the marketplace accounted for differences in booking rates (airbnb price vs expedia booking rates). Due to how the data was structured we saw that the booking behavior was misleading because if a user takes a while to find the hotel that they seek, the database will count them as a not-buyer multiple times and a buyer only once. since we had to aggregate the data we accounted for this by choosing the max variable on the is_booking variable aggregated by userid and checkin date searched. I also created a count variable of the is_booking to be able to repopulate lost information if we desired later. Before we went ahead and tried to train any models we wanted to see what sort of variance the aggregated data accounted for by graphing the means of the count(is_booking) variable for Los Angeles, which we also had airbnb data for. We immediately saw that there was a discrepancy between those that booked and didn't book in Los Angeles and that this discrepancy could be called "pickiness" -- that a user looking to make a purchase will be more picky than a user less invested in making a purchase. Our next hypothesis was that users in cities with different market characteristics would be more or less picky depending on what other options existed in the marketplace. For the first three cites that we tested, this pattern matched the market characteristics exactly. There was one problem. We only tested three cities and they all fell along the lines of larger to smaller cities. We needed a small city with airbnb market characteristics like NYC to validate our hypothesis. This city was Seattle. Seattle debunked our hypothesis. Not totally debunked it, for our new hypothesis was that airbnb was still a source of variance on the expedia data, but what we call "vacationiness"--a cities vacation appeal -- was the larger source of variance on pickiness, for Seattle followed that trend. We doubled the cities tested and they all followed the vacationiness trend that we would expect. **This is where we dropped the ball.** Half of the team was already invested in the airbnb search for variance and continued to search for predictive power between the two data sets, meanwhile we found what we think was a major source of variance and got around to training the random forest late. The random forest model pointed out to us that pickiness was actually the leading source of variance! But the random forest couldn't distinguish vacationiness as a source of variance on is_booking. We should have trained it on the pickiness variable instead and delved into how to model vacationiness since by all appearances the abstract idea of vacationiness is a predictor of pickiness which is a predictor of is_booking.

Thoughts on our presentation and datafest:

If we could do it again I think we should have spent more time making even cleaner and more aesthetic graphics and dropped the airbnb data analysis all together except for in explaining how we found the pickiness variable. We should have mentioned that we dropped user_id from our training model since user_id was a huge predictor of booking behavior, but misled the predictive power of the model on new users, so while the 80% testing accuracy wasn't as impressive as other teams, we only used a subset of the data. A good machine learning algorithm should be able to pick up on the fact that the same userid exists in k rows and has the equivalence of a c*k pickiness score for some constant c. But if we could do it again we also could spend more time searching for a quantitive model to define vacationiness.