Reverse Engineering Data Secrets
Time Delay Booking Predictions

Team Pandas
Ryan Cox, Kristina Yamkovoy, Anubhab Haldar, Hanfei Zhang, Ojaswin Karthikeyan

Reverse Engineering Data Secrets

Slide 1:

We began our exploration of the data by focusing on the destination data, which contains a latitude, longitude pair for each unique destination search ID. We created a heatmap of this data by plotting each destination according to our lat/long pairs and overlayed this on a Google map. We can see most popular search ID's by looking at the density of the points on the map.

The destinations file also has a mysterious 'destination_type_id' variable, containing seven discrete destination type categories. The data dictionary offers no insight to these categories, so we wondered what they could mean. We found this question posted to the Expedia question document and a representative cryptically replied that the IDs 'follow a pattern' and nothing more. Curiosity taking the better of us, we wanted to find what this pattern was.

Our first approach was to plot the destinations colored by type ID; however, the types seemed to be about evenly distributed, so we needed to go a little deeper.

Slide 2:

In an attempt to determine what these categories may be, we decided to take a machine learning approach for our reverse engineering problem. The general idea was to reduce our ~40k x 140 matrix using a Linear Discriminant Analysis on the srch_destination_type_ids column and use this new reduced matrix to classify a test set using a Nearest Neighbors algorithm.

We ended up with a prediction accuracy of ~0.58 for k=~85. At about 60% accuracy, our result is definitely not as accurate as we'd like it to be, but considering a random classifier would have E(Y) ~= 0.15, we are pretty happy with these results.

While this result is far from perfect, we can still use it to then pull out the columns from our LDA with the highest overall weighting and consider them the "biggest indicators" of belonging within one of our categories. When we did this we found that 'popular_activity_snorkling', 'popular_activity_windsurfing', 'popular_activity_ecotourism', 'popular-activity_indoorskiing', 'popular_cultural_entertainment' were the columns that most contributed to overall classification. While, at first these results may seem strange, we can notice that these results seemingly indicate the type of vacation people are going on. Snorkeling and windsurfing are popular activities for a tropical getaway vacation while things like popular cultural entertainment suggests something closer to a sightseeing trip in Europe.

Time Delay Booking Predictions

Slide 3:

The other focus of our investigation was to attempt to predict whether or not a user will actually book a trip with Expedia based primarily on the relation between browse time and searched check-in time and the length of stay. Anubhab, a frequent-flying member of our team, hypothesized that trips of longer duration (and typically distance) are planned out further ahead. Flight tickets for a 13-hour flight are booked months in advance. A trip between two nearby cities can be planned in less than a week.

Using a simple KNN (sklearn.neighbors.KNeighborsClassifier implementation), we achieved a score of ~0.92, showing that we can predict to a reasonable accuracy whether a user will book or not, by using stay duration, booking delay (time between booking and checking in), destination distance, and time of day when the user visits the website.