Name: Michael Hu and David Zhao **Course:** ORIE 4741

AirBnb Destination Prediction

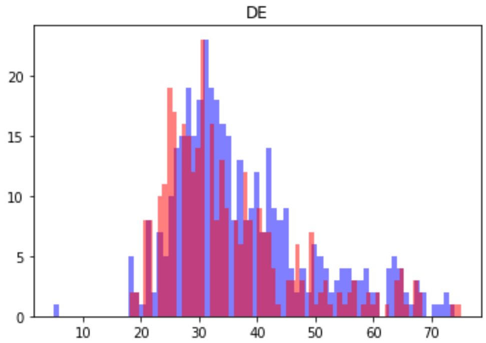
Introduction

With the rise of Airbnb, finding places to stay during trips has never been easier, cheaper, or more centralized. We are interested in determining if the demographics, search histories, and other user data available on Airbnb can be used to accurately predict their next booking location by country. This idea originates from a kaggle competition hosted in 2014, but we intend to build onto the original drive of the competition. Possessing the ability to accurately predict where a user wants to book their next trip allows travel companies to provide targeted ads related to the predicted location, theoretically improving click through rates. As an example, if we know a user lives in the US and is likely intending to travel outside the US, we could give them ads selling European outlet adaptors. Also, proactively knowing when surges in travelers to certain destinations will occur may give companies like Airbnb reason to incentivize more hosts to rent out their spaces in the destinations in question so there is enough space available for everyone.

Data Exploration

Our data is dispersed over four spreadsheets covering user account data, user session data, destination demographics, and destination features. A significant portion of our project will be dedicated towards mining features of interest from the given raw data like the number of searches a user makes or some nonlinear heuristics calculated using a user's and each country's demographic information.

Exploring the user account data, we are given data on each user's age, gender, destination, device type, language, last booking, and account age. Nearly half of the age and gender columns are empty or unreported. Additionally, many of the destinations were undefined or not in the set of countries we were interested in, so those rows had to be dropped. We decided to represent gender as a one-hot encoding of [male, female] where users who didn't decide to identify with either got 0's in both columns. For the missing age data, we intend to test two strategies via cross validation: dropping out users who didn't report an age, or setting each missing age to the average user age. To see how age and gender may affect destination, we plotted histograms of each. Figure 1 shows some interesting trends we observed in our data.



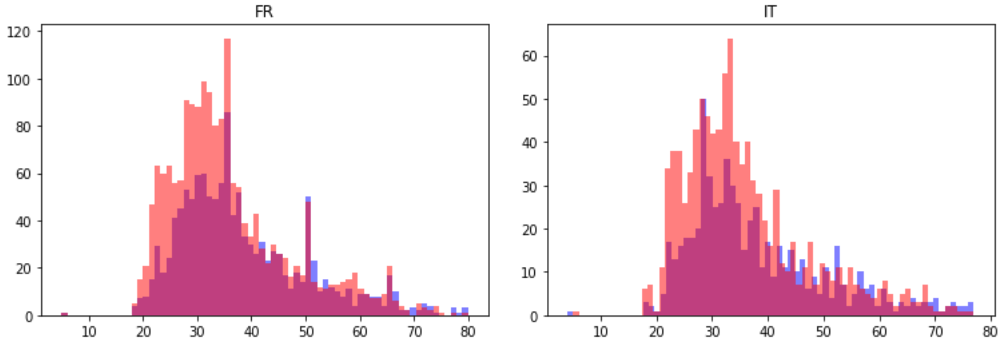


Figure 1: It appears that travelers to Germany (DE) who are older than 30 tend to be male. Travelers to France (FR) under 35 tend to be female, and travelers to Italy (IT) are mostly female. The majority of travelers appear to be 25 to 35 years in age.

In addition to age and gender, we played with ideas for how the other user data could be used. The devices used by users were fairly evenly split, with 37k OSX users and 26k Windows sessions. There are a few iOS and Android data points as well, from mobile sessions. Looking at web session data, we see most users are split between Macs, Windows, and iPhones.

We will try to include device as a feature under the hypothesis that people who own Apple devices may be more willing to spend money and would therefore be more likely to go to more expensive destinations. Language has a less even split, with 97% of users in our dataset speaking English. The second and third most common languages were Chinese and French, each accounting for around .5% of users. Using language as a feature may be useful if there is a trend like non-English speakers generally visiting the US.

Looking at the web session data, what immediately stood out was that there were a lot of recorded actions for different users, a total of 359 different actions (without even looking at the specific types of those actions). These actions cover a wide spectrum of user interactions from searches to updating account info to email validation. Not all of these actions are relevant to where a user booking their first trip, but it is not immediately obvious what actions are worth considering. The top 19 actions cover about 83% of all user interactions as shown in Figure 2.

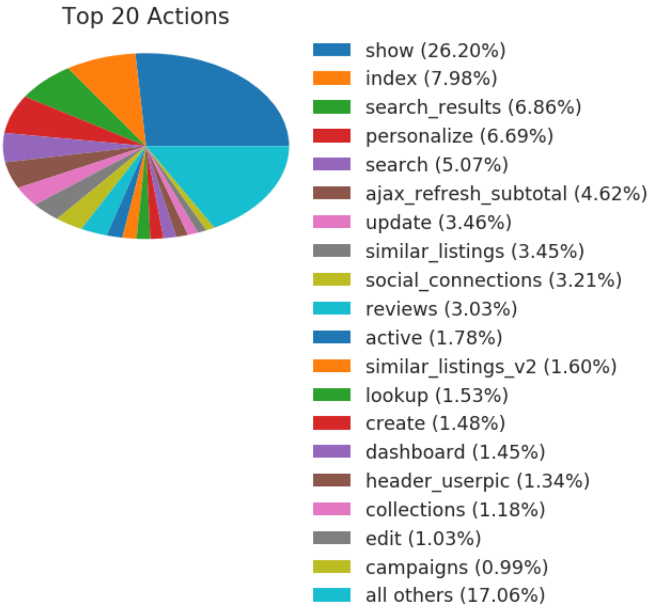


Figure 2. There is a wide range of user interactions, but which ones should we consider? Is there an objective way to approach choosing a selection of actions to consider?

Underfitting and Overfitting

With over 213,000 users data in our training set, we believe that removing data with null values (empty fields or "-unknown-") will still leave us with a large and diverse enough training set. We plan on building our initial models upon the variables we looked at in our preliminary analysis, since a simpler model is always preferable and helps to avoid overfitting to our training data. If we are suspicious of underfitting, we can add back some data points that contained null values by approximate their null values through means (or other measures). Other attempts would include incorporating more features into the model from the data or through use of polynomial kernels. Alternatively, if we are suspicious of overfitting we will rely on L1 regularization to trim our models.

Preliminary Analysis

Just to see how well age and gender predict destination, we built a simple SVM classifier. The training accuracy using either l1 and l2 regularization was around 79.7%. Considering 79.1% of the destinations are the US, these two features aren't enough to build a good classifier.

Testing Metrics

Initially we will divide up our given training data into three different sets: training, validation, and test. For each model we train, we will perform n-fold cross-validation with various folds or simply run the model on our validation set. Considering how 79.1% of the destinations in our dataset are the US, accuracy may not be the best measure of our model. We intend to explore confusion matrices and false positives/negatives in general to measure our error. Our final results will come from how the best model for the validation set performs on the test set.

Moving Forward

Moving forward, we have a lot of data mining and feature engineering ahead of us, as we believe there's still a lot of useful information we can extract from the 359 different actions in the session data and how we interpret them. There is also the question of how we will use the demographic information of each country effectively. We also intend to explore ideas such as building classifiers to detect if a user will likely visit the US or Europe, as that would make better use of our data that is heavily biased towards US destinations.