**Improving Booking Rates – Team LUCAS**

**Data Analysis, Feature Construction and Model Construction:**

We recommend a solution for increasing the number of bookings made on Expedia hotels. First, we determined what constitutes a booking by defining a “user session,” during which a user either ends up booking a hotel or does not. We combined data row entries to create these sessions and labeled each one with our binary classifier, “booked.” In combining these sessions we used the “cnt” column to create the “clicks” feature, which is the number of user clicks per session. We later removed many of the features considered redundant by the model, such as longitude and latitude, and we constructed new features: is\_holiday, stay\_length, and days\_from\_booking (days between hotel booking and site session). Our model confirms that our features were the 2nd, 3rd and 4th most important features for the model.

In order to modify the data, we made certain assumptions. We recognize that these can affect our analyses in potentially significant ways:

1) A session lasts one day. If a user visits the site multiple times in one day, and ends up booking, then that counts as one success (booked=1). If a user visits the site one or more times in a day and does not book, this counts as a failure. We understand this does not cover all corner cases, but the length of a session can tuned with more user information.

2) We deleted cases of null data or incorrectly entered data (year=20015, check in date after check out date)  values in our selected features.

In order to extract a low dimensional set of features from a high dimensional data set we conducted principal component analysis by converting the 4 categorical variables into numeric using one hot encoding. With 39 principal components loading, approximately 34 components explain around 98.9% variance in the data set.

We used our constructed training set to train a gradient boosted trees model with a binary logistic objective function. This model outputs the probability of whether a new user session will result in a booking. With ~3 million observations we trained on a 75% training set and 25% test set. Using 0.5 as our classification threshold on the probability output we reached a MSE of 15.9%, translating to an accuracy of 84.1%, and an AUC of 78.7%. We confirmed our model’s superior performance to simpler models by training the same dataset on logistic regression and random forests, both of which performed at MSE’s of ~20%.

**Actionable Recommendation:**

We recommend the following strategy to improve user bookings and increase revenue.

* As a user inputs session data, the propensity model calculates the probability they will book a hotel, based on the feature values they display.
* Then, as the user accumulates clicks in their session, if they pass the threshold of clicks we calculated (the average number of clicks in sessions that resulted in a booking), the Expedia site will react in one of the following ways:
  + 1) If the model predicts they are highly likely to book, the site offers a live-chat pop-up that offers assistance and answers questions because the user may be confused or unsure of a particular hotel.
  + 2) If the user is predicted to be unlikely to book, but is still displaying interest in the hotels because they have clicked more times than the threshold, the site will pop up offer a discount or coupon with a comparison of the hotel price to Expedia’s competitor prices.

In addition to our insight, we suggest Expedia add the following features to the model: number\_of\_children, and distance\_from\_popular\_attractions. This will aid in clustering user sessions into specific categories of trips, such as a business trip or a family vacation. This will aid in curating the hotels we display to the user. We attempted k-means clustering of the user sessions to determine characteristics of the session bookings, but we were unable to derive any useful insights without additional data.