Summer Wang

Yawen Ye

Matt Staton

Wenduo Wang

**Project goals**

In this project, we investigated a hotel booking dataset randomly selected by Expedia Inc. from the company's database. According to Expedia Inc. “Currently, Expedia uses search parameters to adjust their hotel recommendations, but there aren't enough customer specific data to personalize them for each user. In this competition, Expedia is challenging Kagglers to contextualize customer data and predict the likelihood a user will stay at 100 different hotel groups.” Here we generally hold purpose, and further utilize the data to derive insights into the company's business situation and user base.

**Prospect of the underlying solution**

A common theme of applying machine learning and predictive technology across business spectrum is to provide a higher level of customized services. It is no exception in this study. Expedia (NASDAQ: EXPE) is an international online booking conglomerate, whose customers are mostly individuals with diverse demands, and therefore offering indiscriminate products poses considerable risks to its business. As consumers have complained about annoying promotions when searching for a hotel, the idea of the website knowing users' preferences could create a win-win situation. As to be demonstrated later in this paper, with a deeper understanding of user browsing history, the website essentially narrows its content to what the user cares about, which improves consumption satisfaction and encourages recurring orders. The above benefit can be generalized to a broader audience demanding similar services not limited to travelers.

A bonus for the company, moreover, is the enhanced efficiency with which the data is handled. As pointed out in Expedia Inc.'s 2014 annual report, data center related cost in 2013 and 2014 totaled 199 and 229 million dollars respectively, accounting for 4% of total revenue and 61%, 92% of the annual net income. The potential profit improvement through leaner data analytics provides extra motivation for the business to employ novel technologies.

**A peek into the data**

The provided data consists of 3 parts: a training set consisting of ~37million rows, a test set of ~2.5million rows, and a separate destinations file containing attribute decomposition for all destinations.

As stated at the beginning, the goal is to predict the probabilities of a user booking different hotel groups, and in the datasets, hotel groups are coded as ***hotel\_cluster***.

In the training set, the time span is from 2013 to 2014, and there are 24 columns recording many users' information and browsing history, including the hotel clusters the user viewed. If the user booked the hotel, the column ***is\_booking*** flags 1.

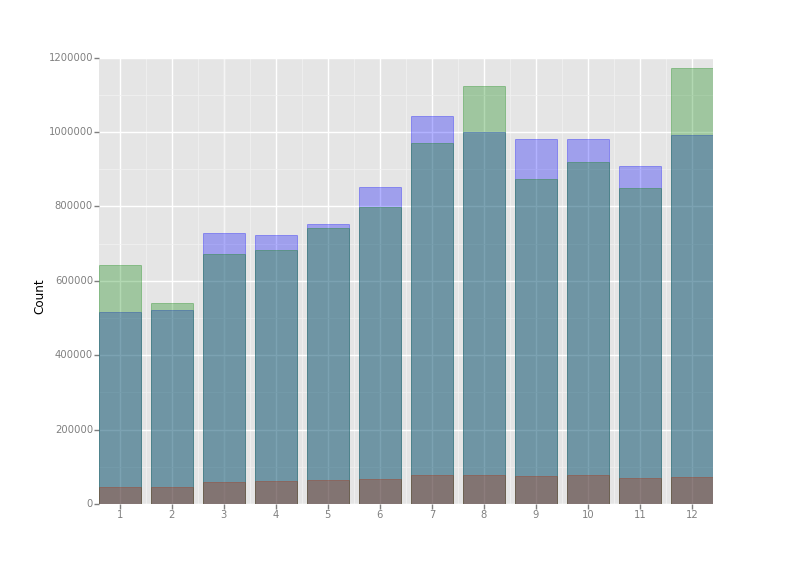
The records in the test set are from 2015, and therein ***hotel\_cluster*** is left out for prediction, while ***is\_booking*** and browsing history count ***cnt*** are also taken out, since all records are successful transactions in the test set, so by default ***is\_booking*** is 1. All ***user\_id*** have appeared in the training set.

The destinations file lists 150 attributes for each destination in the training set. This file is intended to simplify the ***srch\_destination\_id*** variable in the training and test sets, for the ***srch\_destination\_id*** contains more than 60 thousand distinct values.

**Intuitions on the data**

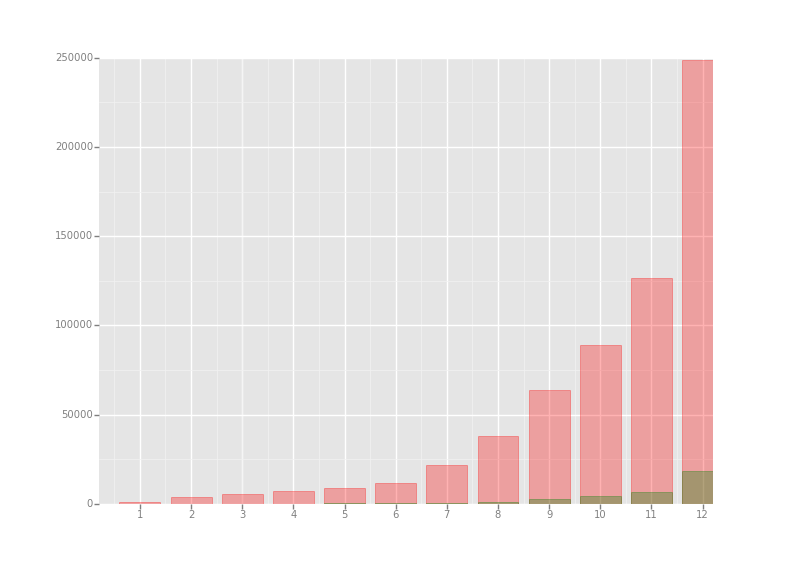
The abundance of data enables convenient exploration into different aspects of Expedia's business model and user base, before diving into quantitative modeling.

An important popularity indicator of a website is monthly volume, so let's make a bar graph of monthly total click throughs to evaluate it. Here we regard all records as from the same portal since the data is merged together, yet it should be noticed that there are many sites under the Expedia brand.

Illustration 1: Monthly click thrus (blue) including actual booking (darkest) and check-in month (green)

The graph shows a clear seasonal pattern, and the volume in the first quarter as a whole is relatively low. Yet it is interesting to compare this phenomenon with the MD&A section of Expedia's annual report which states “... *traditional leisure travel bookings are generally the highest in the first three quarters as travelers plan and book their spring, summer and holiday travel. The number of bookings typically decreases in the fourth quarter. Because revenue for most of our travel products, including merchant and agency hotel, is recognized when the travel takes place rather than when it is booked, revenue typically lags bookings by several weeks or longer.*”

Let's have a closer look at, say, when people search for hotels for Christmas and summer holidays, and when they actually place an order.

Illustration 2: Monthly search for Christmas hotels, actual bookings in green.

Another way to understand the company's business is to look at its user activity and recurring visits. Plotting a bar graph on ***user\_id*** counts gives insight into this aspect.