Group P.H.A.S.S Project Summary
DataFest 2017

Angie Shen, Sam Yin, Haozhang Jiang, Pim Chuaylua

May 3, 2017

1 Introduction

For our project, we looked at the times of search - the number of searches from the same user before finally booking with a continuation ratio probit model, and the exact time of search - to compare with the New York Times hits, a metric we chose to represent possible outside influences with respe

2 Continuation Ratio Probit Model

For each trip, a typical user of Expedia.com searched several times before he/she eventually booked the hotel or sought other options. If we consider the decision of booking as an event and each preceding search as a discrete time point to the event, the probability of user i booked a hotel provided that he/she didn't book in the previous searches and other covariates can be modeled as

$$\Phi^{-1}(\Pr[T_i = t | T_i \geq t, \mathbf{X}_i]) = \alpha_t + \mathbf{X}_i^T \boldsymbol{\theta}$$

where T_i is the number of times that user i has searched until they booked, $\mathbf{X}_i$ is the set of covariates including: the median distance between the hotels searched and the user, the indicator for mobile device connection, the indicator for package search, the channel ID, the number of adults, children, and hotel rooms specified in the search, the indicator for major hotel chain, the star rating of the hotel, banded distance of the hotel relative to other hotels in the same destination, banded historical purchase price of a hotel relative to other hotels in the same destination, and banded hotel popularity relative to other hotels in the same destination. Φ^{-1} is the inverse CDF of a standard normal distribution as a link function, $\Phi(\alpha_t)$ is the baseline conditional probability of booking at t^{th} search, and $\boldsymbol{\theta}$ is the vector of coefficients describing the shift in the expected conditional probability of booking induced by each unit of change in the covariate.

The model was fit to data for users located in the United States of America and that for Germany and compare the estimates of the conditional probability and covariates. We have seen a similar pattern of covariates' effect, but the American users were more likely to book on the second or third searches while the German users had a consistent chance of booking throughout all searches. The difference in the time required for a booking decision can inform the type of advertisement and recommendation display tailored to users from specific countries.

The current model does not consider time-variant covariates, which are likely to affect the baseline probability. The inference could also be based on posterior samples from a Markov Chain Monte Carlo algorithm with data augmentation.

3 Outside Data: New York Times Articles

For this part of the analysis, we used New York Times articles as indications of the social network "buzz" around any particular location. We used a time interval of a week, from Monday to Sunday, to divide the data into 51 periods. For the data that we're looking at, in week 46 of 2015, there was the terror attack in Paris (on November 13th), that corresponds to a very steep increase in the number of hits of New York Times articles, from 698 to 1296. At the same time, there is a very steep drop in hotel searching behavior on Expedia, from 1876 in week 45 to 1146 in week 46. As there is no data from previous years, we are not able to compare the result with a baseline, but the number of searches from the holiday season (around 1000) is even lower than in the beginning of the year (around 1500).

The process of how we obtained the data from New York Times is described as follows. We used the New York Times Developer API to get the number of views of articles with the key word mentioned within each week. The script is written in PHP.

4 Discussion

We initially tried to model the data as a network: users in Country A sending ties to hotels in Country B, and vice versa. Then we might be able to too run a "network" regression to determine what kind of covariates is more explanatory in the booking behavior of clients, and are those different for users and hotels in different areas. The problem with this kind of modeling is that the data we're working with is more about the user specification and less about the socioeconomic factors that might play a huge role in such interactions.