Project

Choice Modeling, Assortment Optimization and Pricing

- Due date: Thursday April 27th at 11:59pm (Eastern Time) on Gradescope.
- Late submissions will NOT be accepted.
- Each individual must submit the group report.
- For any question where you use `Python`, `R` or `java`, submit your code and output.

Data

The dataset we will be working with is provided by Expedia as part of a Kaggle competition [1]. It gives the results of search queries for hotels on Expedia. The dataset `data.csv` can be downloaded from Canvas.

The rows of the dataset correspond to different hotels that are displayed in different search queries. The columns give information on the characteristics of the displayed hotels in a search query and the booking decision of the customer. The dataset contains 153,009 rows which represent 8354 queries.

In all the queries, the customer is looking for a single night booking. Each row corresponds to a displayed hotel in a search query. The columns in the dataset have the following interpretation: The first column is the unique code of the search query in which the hotel was displayed. Using this column, we have access to all the displayed hotels in the search query. This corresponds to the set of hotels among which the customer needs to make a choice. The last column is an indicator of whether the customer booked the hotel in the search query. This corresponds to the purchase decision of the customer. A customer can book at most one hotel or leave without making any booking.

These eight columns show the attributes of the displayed hotel: Star rating, review score, hotel brand, location score, accessibility score, historical price, displayed price and promotion flag. The following five columns give information about the customer making the search query: Booking window of the customer, i.e., the number of days between the time of the search query and the booking date that the customer is looking for, the number of adults in the intended booking, the number of children in the intended booking, the number of rooms the customer is looking for and lastly an indicator of whether the customer is looking for a Saturday night booking. A snapshot of the data is presented in Table 1.

Search ID	Star rating	Review score	Hotel brand	Location	Accessibility	Log historical price
360623	3.0	4.0	1.0	0.0	0.0	5.0
360623	3.0	5.0	1.0	1.0	0.0	5.0

Displayed price	Promotion	Booking window	# Adults	# Children	# Room	Saturday	Booking
150.0	0.0	25.0	1.0	2.0	1.0	1.0	0
140.0	0.0	25.0	1.0	2.0	1.0	1.0	1

Table 1: Example of the data.

[1] `https://www.kaggle.com/c/expedia-personalized-sort`

Data for this Homework

For this homework, we will be using the full dataset `data.csv` for estimating choice models. There are also four small datasets `data1.csv`, `data2.csv`, `data3.csv` and `data4.csv` provided. These four datasets contain features for a list of different hotels, and we will use them for assortment optimization and pricing problems.

Problem 1: MNL Model

Assume customers make choices according to an MNL model. Let v_j be the preference weight of hotel j and define it as

$$v_j = e^{u_j}, \;\; u_j = \beta_0 + \sum_{i=1}^{8} \beta_i x_{ji},$$

where $x_{ji}, i = 1, \ldots, 8$ are the features of hotel j and β_i is the sensitivity of customer to feature i. The probability of customer choosing hotel j given a set of displayed hotels S is

$$\mathbb{P}(j|S) = \frac{v_j}{1 + \sum_{p \in S} v_p}.$$

Use star rating, review score, hotel brand, location score, accessibility score, historical price, displayed price and promotion flag as the features of hotels. Estimate the parameters $\beta_0, \ldots, \beta_8$ using MLE estimation. Comment on the coefficient of each of the features.

Problem 2: Assortment Optimization under MNL

Assume customers make choices according to the MNL model we estimated in Problem 1. Given the set of hotels in `data1.csv`, suppose you want to show a subset of these hotels to the customers, what is the optimal subset of hotels to display? Give the expected revenue under this optimal assortment. Repeat the same question for `data2.csv`, `data3.csv` and `data4.csv`.

Problem 3: Pricing under MNL

Assume customers make choices according to the MNL model we estimated in Problem 1. Consider the set of hotels in `data1.csv`. Suppose the company will display all these hotels to customers. The company wants to change the price of each of these hotels (column price_usd in the data). What are the optimal hotel prices that maximizes the expected revenue under MNL model given that we display all of them. Repeat the same question for `data2.csv`, `data3.csv` and `data4.csv`.

Problem 4: Mixture of MNL

Define two types of customers based on whether the customer wants to make an *early* or a *late* reservation. Consider the column *Booking window*, if the booking window is less than 7 days, we define it as a late booking and early otherwise.

Let θ_1 be the probability that a customer belongs to type 1 (early customers) and θ_2 be the probability that a customer belong to type 2 (late customers).

Let v_{jk} denotes the preference weight of hotel j for customer type k, and

$$v_{jk} = e^{u_{jk}}, \;\; u_{jk} = \beta_{0k} + \sum_{i=1}^{8} \beta_{ik} x_{ji},$$

where β_{ik} is the sensitivity of type k customers to feature i. The probability of customer choosing hotel j given a set of hotels S is

$$\mathbb{P}(j|S) = \sum_{k=1}^{2} \theta_k \frac{v_{jk}}{1 + \sum_{p \in S} v_{pk}}.$$

Use the full dataset `data.csv` to estimate θ_1 and θ_2 by computing the size of customers of each type. We will estimate an MNL model for each type. Use star rating, review score, hotel brand, location score, accessibility score, historical price, displayed price and promotion flag as the features of hotels. Estimate the sensitivity parameters for each type of customers using MLE estimation. Comment on the difference in the sensitivity parameters of these two types of customers.

Problem 5: Early vs. Late Reservations

Assume customers make choices according to the mixture of MNL model we estimated in Problem 4. Given the set of hotels in `data1.csv`, suppose you want to show a subset of these hotels to the customers that maximizes the revenue of the company.

- Assume we don't know the type of an arriving customer. What is the optimal subset of hotels to display? Let's call it S. You need to solve an integer program here to compute S.
- Suppose we know that the arriving customer is of type 1. What is the optimal subset of hotels to display? Let's call it S_1.
- Compare the expected revenue of S and S_1 under MNL model of type 1.
- Suppose we know that the arriving customer is of type 2. What is the optimal subset of hotels to display? Let's call it S_2.
- Compare the expected revenue of S and S_2 under MNL model of type 2.

Repeat the same questions for `data2.csv`, `data3.csv` and `data4.csv`.

Problem 6: Mixture of MNL with Other Ways of Defining Customer Types

In previous problems, we define customer types based on whether the customer wants to make an early or a late reservation. Given dataset `data.csv`, there are other ways to define customer types, for example, whether the customer in the search query is looking for a Saturday night booking. Explore your own ways to define customer types and estimate the mixture of MNL model. Repeat Problem 5 for this new MMNL model.