Udacity Machine Learning Nanodegree

Expedia Hotel Recommendations (Which hotel type will an Expedia customer book?)

Rajat Hore

March 20th, 2019

Project Overview:

When we are planning for a dream vacation, or even a weekend escape, can be an overwhelming affair. With hundreds, even thousands, of hotels to choose from at every destination, it's difficult to know which will suit our personal preferences. Should we go with an old standby with those pillow mints you like, or risk a new hotel with a trendy pool bar?

Expedia wants to take the proverbial rabbit hole out of hotel search by providing personalized hotel recommendations to their users. This is no small task for a site with hundreds of millions of visitors every month!

Currently, Expedia uses search parameters to adjust their hotel recommendations, but there aren't enough customer specific data to personalize them for each user. The challenge here is to contextualize customer data and predict the likelihood a user will stay at 100 different hotel groups.

Problem Statement:

Expedia has provided logs of customer behavior. These include what customers searched for, how they interacted with search results (click/book), whether or not the search result was a travel package. The data is a random selection from Expedia and is not representative of the overall statistics.

We need to predict which hotel group a user is going to book. Expedia has in-house algorithms to form hotel clusters, where similar hotels for a search (based on historical price, customer star ratings, geographical locations relative to city center etc.) are grouped together. These hotel clusters serve as good identifiers to which types of hotels people are going to book, while avoiding outliers such as new hotels that don't have historical data.

My goal is to predict the booking outcome (hotel cluster) for a user event, based on their search and other attributes associated with that user event.

Data Exploration:

The **train and test datasets** are split based on time: training data from 2013 and 2014, while test data are from 2015. The public/private leaderboard data are split base on time as well. Training data includes all the users in the logs, including both click events and booking events. Test data only includes booking events.
destinations.csv data consists of features extracted from hotel reviews text.
Note that some srch_destination_id's in the train/test files don't exist in the destinations.csv file. This is because some hotels are new and don't have enough features in the latent space. Your algorithm should be able to handle this missing information.

File descriptions

- **train.csv - the training set**
- **test.csv - the test set**
- **destinations.csv - hotel search latent attributes**
-

Data fields

- train/test.csv----------

Column name	Description
date_time	Timestamp
site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)
posa_continent	ID of continent associated with site_name
user_location_country	The ID of the country the customer is located
user_location_region	The ID of the region the customer is located
user_location_city	The ID of the city the customer is located
orig_destination_distance	Physical distance between a hotel and a customer at the time of search. A null means the d could not be calculated
user_id	ID of user

Column name	Description
is_mobile	1 when a user connected from a mobile device, 0 otherwise
is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise
channel	ID of a marketing channel
srch_ci	Checkin date
srch_co	Checkout date
srch_adults_cnt	The number of adults specified in the hotel room
srch_children_cnt	The number of (extra occupancy) children specified in the hotel room
srch_rm_cnt	The number of hotel rooms specified in the search
srch_destination_id	ID of the destination where the hotel search was performed
srch_destination_type_id	Type of destination
hotel_continent	Hotel continent
hotel_country	Hotel country
hotel_market	Hotel market
is_booking	1 if a booking, 0 if a click
cnt	Numer of similar events in the context of the same user session
hotel_cluster	ID of a hotel cluster

- Destinations.csv----------

Column name	Description
srch_destination_id	ID of the destination where the hotel search was performed
d1-d149	latent description of search regions

Algorithms and Techniques:

I am going to use the below list of algorithms on the given dataset and will check which performs better among them

I will use PCA to reduce the size of the dataset while preserving the variance between rows

Algorithms to evaluate----------
- Logistic Regression (LR)
- Support Vector Machine
- Random Forests (RF)
- KNeighbors Classifier
- Naïve Bayes

Evaluation Metrics:

Evaluation metrics will be calculated according to the Mean Average Precision @ 5 (MAP@5):

$$MAP@5 = \frac{1}{|U|}\sum_{u=1}^{|U|}\sum_{k=1}^{min(5,n)} P(k)$$

Where |U| is the number of user events, P(k) is the precision at cutoff k, n is the number of predicted hotel clusters.

Output File:

For every user event, we must predict a space-delimited list of the hotel clusters they booked. We need to submit up to 5 predictions for each user event. The file should contain a header and have the following format:

```
id, hotel_cluster
0,99 3 1 75 20
1,2 50 30 23 9
etc...
```

Project Design:

The workflow of solving this problem will be in the following order:

- **Exploring the Data**
  - Loading Libraries and data
  - Peek at the training data
  - Dimensions of data
  - Overview of responses and overall response rate
  - Statistical summary

- **Data preprocessing/cleaning**
  - Preprocess feature columns
  - Identify Feature and Target columns
  - Data cleaning
  - Training and Validation data split
  - Feature Scaling - Standardization/Normalizing data

- **Evaluate Algorithms**
  - use various algorithms on the dataset
  - Select best algorithms(model) as per Mean Average Precision

- **Model Building and then Tuning to Improve Result**

- **Final conclusion**

  - Need to submit up to 5 predictions for each user event. The output file structure will be like:

```
id, hotel_cluster
0,99 3 1 75 20
1,2 50 30 23 9
etc...
```

Benchmark model:

I will see which algorithms gives the best result on the given dataset and accordingly I will try to boost it for better accurate result.

References:

https://www.kaggle.com/c/expedia-hotel-recommendations/data