VU Data Mining 2016 - Assignment 2

Mark Hoogendoorn and Ward van Breda
VU University Amsterdam

Start of assignment: April 17, 2015
Deadline: May 29, 2016 (23:59)

Abstract. This document describes Assignment 2 of the Data Mining Techniques course of 2015. Please make sure you read it thoroughly and carefully. This is a group task (maximum 3 members), and please make sure all team members contribute to the work as expected. There will be three things to be submitted: 1) A report about the results; 2) A prediction file, and (3) A process report. For details, please see below.

1 Motivation

By now should have a fair idea about techniques we can use, and should also have some practical experience with mining datasets. In this second assignment, you will gain more experience, explore various techniques (and whether they work in this situation), and hopefully learn a lot. The topic of this assignment is positioned in the area of recommender systems. More specifically, your task is to predict what hotel a user is most likely to book. This could greatly help companies such as Expedia (from which the dataset actually originates) to organize the search results for a user in the most suitable way.

2 Dataset

The dataset, which can be downloaded from BB and originates from a Kaggle competition[1], is split into a training and a test set (train.csv and test.csv respectively, each containing approximately 5 million records). Essentially, the dataset contains information about a search query of a user for a hotel, the hotel properties that resulted and for the training set, whether the user clicked on the hotel and booked it. More in specific, the following fields are present[2]:

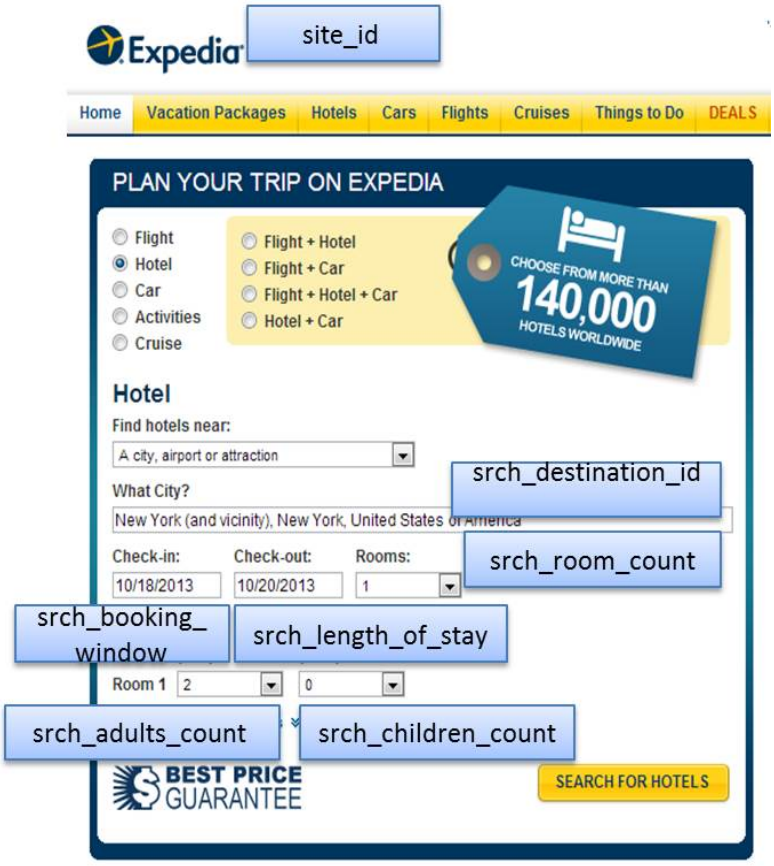
Field	Data Type	Description
srch_id	Integer	The ID of the search
date_time	Date/time	Date and time of the search
site_id	Integer	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ..)
visitor_location_country_id	Integer	The ID of the country the customer is located
visitor_hist_starrating	Float	The mean star rating of hotels the customer has previously purchased; null signifies there is no purchase history on the customer
visitor_hist_adr_usd	Float	The mean price per night (in US$) of the hotels the customer has previously purchased; null signifies there is no purchase history on the customer
prop_country_id	Integer	The ID of the country the hotel is located in
prop_id	Integer	The ID of the hotel
prop_starrating	Integer	The star rating of the hotel, from 1 to 5, in increments of 1. A 0 indicates the property has no stars, the star rating is not known or cannot be publicized.
prop_review_score	Float	The mean customer review score for the hotel on a scale out of 5, rounded to 0.5 increments. A 0 means there have been no reviews, null that the information is not available.
prop_brand_bool	Integer	+1 if the hotel is part of a major hotel chain; 0 if it is an independent hotel
prop_location_score1	Float	A (first) score outlining the desirability of a hotel's location
prop_location_score2	Float	A (second) score outlining the desirability of the hotel's location
prop_log_historical_price	Float	The logarithm of the mean price of the hotel over the last trading period. A 0 will occur if the hotel was not sold in that period.
price_usd	Float	Displayed price of the hotel for the given search. Note that different countries have different conventions regarding displaying taxes and fees and the value may be per night or for the whole stay
promotion_flag	Integer	+1 if the hotel had a sale price promotion specifically displayed
srch_destination_id	Integer	ID of the destination where the hotel search was performed
srch_length_of_stay	Integer	Number of nights stay that was searched

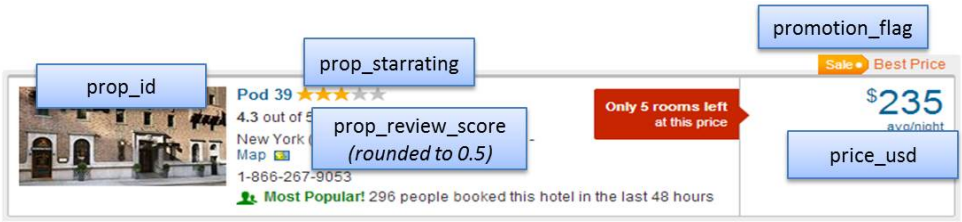
[1] https://www.kaggle.com/c/expedia-personalized-sort or go to Kaggle.com > Competitions > All competitions > Personalize Expedia Hotel Searches ICDM 2013

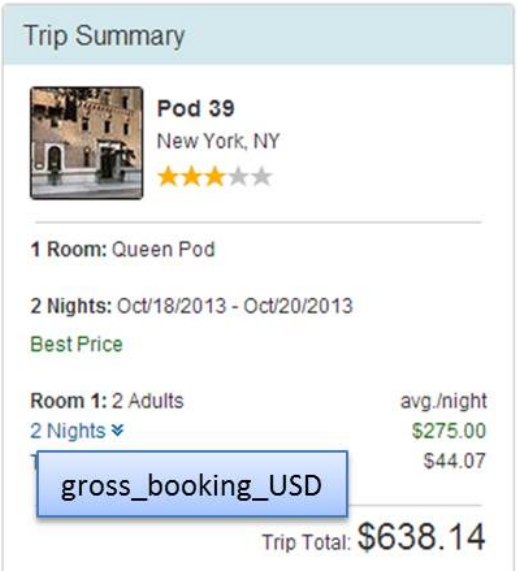
[2] Primarily based on https://www.kaggle.com/c/expedia-personalized-sort/data or follow the description above and select "data" from the menu.

srch_booking_window	Integer	Number of days in the future the hotel stay started from the search date
srch_adults_count	Integer	The number of adults specified in the hotel room
srch_children_count	Integer	The number of (extra occupancy) children specified in the hotel room
srch_room_count	Integer	Number of hotel rooms specified in the search
srch_saturday_night_bool	Boolean	+1 if the stay includes a Saturday night, starts from Thursday with a length of stay is less than or equal to 4 nights (i.e. weekend); otherwise 0
srch_query_affinity_score	Float	The log of the probability a hotel will be clicked on in Internet searches (hence the values are negative) A null signifies there are no data (i.e. hotel did not register in any searches)
orig_destination_distance	Float	Physical distance between the hotel and the customer at the time of search. A null means the distance could not be calculated.
random_bool	Boolean	+1 when the displayed sort was random, 0 when the normal sort order was displayed
comp1_rate	Integer	+1 if Expedia has a lower price than competitor 1 for the hotel; 0 if the same; -1 if Expedia's price is higher than competitor 1; null signifies there is no competitive data
comp1_inv	Integer	+1 if competitor 1 does not have availability in the hotel; 0 if both Expedia and competitor 1 have availability; null signifies there is no competitive data
comp1_rate_percent_diff	Float	The absolute percentage difference (if one exists) between Expedia and competitor 1's price (Expedia's price the denominator); null signifies there is no competitive data
comp2_rate		
comp2_inv		(same, for competitor 2 through 8)
comp2_rate_percent_diff		
. . . .		
comp8_rate		
comp8_inv		
comp8_rate_percent_diff		
Training set only		
position	Integer	Hotel position on Expedia's search results page. This is only provided for the training data, but not the test data.
click_bool	Boolean	1 if the user clicked on the property, 0 if not.
booking_bool	Boolean	1 if the user booked the property, 0 if not.
gross_booking_usd	Float	Total value of the transaction. This can differ from the price_usd due to taxes, fees, conventions on multiple day bookings and purchase of a room type other than the one shown in the search

Each line in the dataset represents a combination of a search query by a user with one specific hotel property that was shown as part of the results. Of course, a list of hotels is presented to the user (and hence, there are multiple rules describing a single search). Lines that belong to the same user/search are identified by the same search id. The link between the fields shown above and the Expedia site are shown graphically below.


site_id
Expedia
Home
Vacation Packages
Hotels
Cars
Flights
Cruises
Things to Do
DEALS
PLAN YOUR TRIP ON EXPEDIA
Flight
Hotel
Car
Activities
Cruise
Flight + Hotel
Flight + Car
Flight + Hotel + Car
Hotel + Car
CHOOSE FROM MORE THAN
140,000
HOTELS WORLDWIDE
Hotel
Find hotels near:
A city, airport or attraction
What City?
New York (and vicinity), New York, United States of America
srch_destination_id
Check-in:
Check-out:
Rooms:
10/18/2013
10/20/2013
1
srch_room_count
srch_booking_window
srch_length_of_stay
Room 1
2
0
srch_adults_count
srch_children_count
BEST PRICE GUARANTEE
SEARCH FOR HOTELS


promotion_flag
prop_starrating
prop_id
Sale
Best Price
Pod 39
4.3 out of 5
New York
Map
prop_review_score
(rounded to 0.5)
1-866-267-9053
Most Popular! 296 people booked this hotel in the last 48 hours
Only 5 rooms left at this price
$235
avg/night
price_usd


Trip Summary
Pod 39
New York, NY
1 Room: Queen Pod
2 Nights: Oct/18/2013 - Oct/20/2013
Best Price
Room 1: 2 Adults
avg./night
2 Nights
$275.00
$44.07
gross_booking_USD
Trip Total: $638.14

3 Detailed task description

To make things easier, we'll use a DM process model to describe your task in a bit more detail:

3.1 Business understanding

Your task is to predict what hotels properties listed as a result of a hotel search a user is most likely to click on. Of course, more people have worked on such predictions. Can you find some other people that have tried to make such predictions (e.g. from the Kaggle competition)? And what have they used as most prominent predictors? Have other people that participate in the competition mentioned anything about their approaches? Please spend a couple of paragraphs on this topic (i.e. related work) in your report.

3.2 Data understanding

Essentially, this is a subtask that requires you do exploratory data analysis (EDA). Explore the dataset, count, summarize, plot things, and report findings that are useful for your task. Remember that EDA is not necessary done once and then you move on. It might very well be possible that you do some EDA, build some models, then some idea comes up, do some more EDA, modify your model according to what it shows, and so on.

3.3 Data preparation

You'll certainly need to work on the dataset, to create, modify or add new features. For instance, you might want to compare the different properties that resulted from the search instead of learning from them one by one. There are certain attributes with a large amount of missing values, do they still provide useful information? And how will you handle a missing value if this shows to be the case? Finally, in order to test your approach (since you do not know the answers for the test set) you will need to split up your data to test your approach yourself before you generate your answers on our test set. Of course, you are also allowed to use external data sources if you find ones that are useful.

3.4 Modeling and Evaluation

Naturally, once you prepare the dataset, you should be able to build models. Use whatever technique you want, though the choice might be influenced by how we would like to measure your predictions at the end (see Section 5.1).

3.5 Deployment

Not really deployment, as you are (probably) not building a commercial product, but we'd like be sure that your results are reproducible, so I might select some people (randomly, or by competition rank, I will decide later) and ask them to demonstrate whether and how their approach works.

4 Deliverables and formats

We have covered the process above, let us see what we expect you to deliver.

4.1 Predictions

You'll need to submit a single file, which ranks the properties belonging to a user search on the likeliness that the property will be booked. Here, you should start with listing the property most likely to be booked. An example of part of such a file is shown below.

```
SearchId ,PropertyId
2, 7771
```

2, 26540
2, 25579
2, 7374
2, 131173
2, 37331
2, 27090
2, 12938
2, 78858
2, 30434
2, 91899
2, 3105
2, 6399
3, 130729
3, 103937
3, 55688

4.2 Scientific Report

The assignment is not only about winning, but also about quality of the process and understanding of what you did. Therefore, we would like you to write a report, which should contain the following:

1) What you did (you might want to follow the process model, and describe the steps you took. If you tried a number of things but only some worked, please mention those that did not work as well, and discuss why they might not have worked).
2) What you learned (either inside the main part of the report, or separately in a paragraph of two, please describe what skills and knowledge you have gained from this assignment, what were the main difficulties, expected and unexpected outcomes of your experiments, etc.
3) Please format the report according to the lncs guidelines (exactly the way you used previously when doing assignment 1).

The report should be maximum 10 pages long. Again, it is 10 pages including possible reference lists and appendices.

4.3 Process Report

As the assignment is done in a group, we would like to get insight into what each individual group member contributed to the eventual result. Therefore we ask you to compose a process report of at most 2 pages which addresses:

1) A schedule of when you performed what task (e.g. on April 28 we explored the dataset and looked for suitable approaches for the task at hand).
2) Who contributed to what task (e.g. Mark was responsible for transforming the dataset into a suitable format for SPSS Modeler whereas Berend was working on the report).
3) A critical reflection of the overall cooperation within the team.

5 Evaluation

Here's how you will get rewarded for your work:

5.1 Winning the competition

The winner will be rewarded with the fame and glory of winning the 2015 VU Data Mining Techniques cup. Your accuracy score will be determined as follows (cf. Kaggle):

The evaluation metric for this competition is Normalized Discounted Cumulative Gain calculated per query and averaged over all queries. See
https://www.kaggle.com/wiki/NormalizedDiscountedCumulativeGain
for more details.

Hotels for each user query are assigned relevance grades as follows:

5 - The user purchased a room at this hotel
1 - The user clicked through to see more information on this hotel

0 - The user neither clicked on this hotel nor purchased a room at this hotel

Submissions for each user query should recommend hotels in order from the highest grade (most likely to purchase a hotel room) to the lowest grade (least likely to purchase a hotel room or click on the hotel). We know that the correct values for the test set are available online, of course, you are not allowed to use this. If we suspect that you used those values we will ask you for your code and check whether your results are reproducible using the training set as a basis for generating your predictive model.

5.2 Getting a high grade

You can get 100 marks for this assignment, so half of your final grade depends on this. 60% of this mark can be achieved by submitting a nice and thorough report, 20% will come from where you end up in the competition and 20% will be based on the process report. Regarding the competition-based marks, we will see how good your submissions are, and then we'll decide if you'll get the marks based on accuracy value, or rank in the competition.

6 Submission

The submission consists of two parts: (1) the prediction file, and (2) the written reports. The deadline for both parts is 29 May 2016 (23:59). Both the prediction and the report should be submitted via Blackboard.

7 Closing events

The presentation of your final assignment will be done during the closing event on May 31 2016. Here, we will present the final outcome of the competition (given the predictions you handed in the weekend before) and hand out the cup and the fame and glory to the lucky winners. Each of the groups is expected to prepare a presentation (of *at most 10 minutes*), those groups that ended up in the top 3 and three additional random groups will be asked to present their work.