VRIJE UNIVERSITEIT AMSTERDAM

Assignment 2
Data Mining Techniques
A Real Life Competition

Deadline Competition: 19/05/2024, 23:59
Deadline Report: 24/05/2024, 23:59

INTRODUCTION

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.

This document describes Assignment 2 of the Data Mining Techniques course at the VU. 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 file to be uploaded on the VU DMT Kaggle competition, and (3) A process report.

DATASET AND PROBLEM

The dataset can be downloaded from our in class Kaggle website `https://www.kaggle.com/competitions/dmt-2024-2nd-assignment`. For signing up for the competition, please fol-

low the following link: `https://tinyurl.com/join-dmt-cup`.
The dataset originates from a former Kaggle competition[1]. Using the dataset from the original Kaggle competition is not allowed. The data is split into a training and a test set, train.csv and test.csv 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. The field that are present are shown in Table 1[2].

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 in Figures 1-3[3].

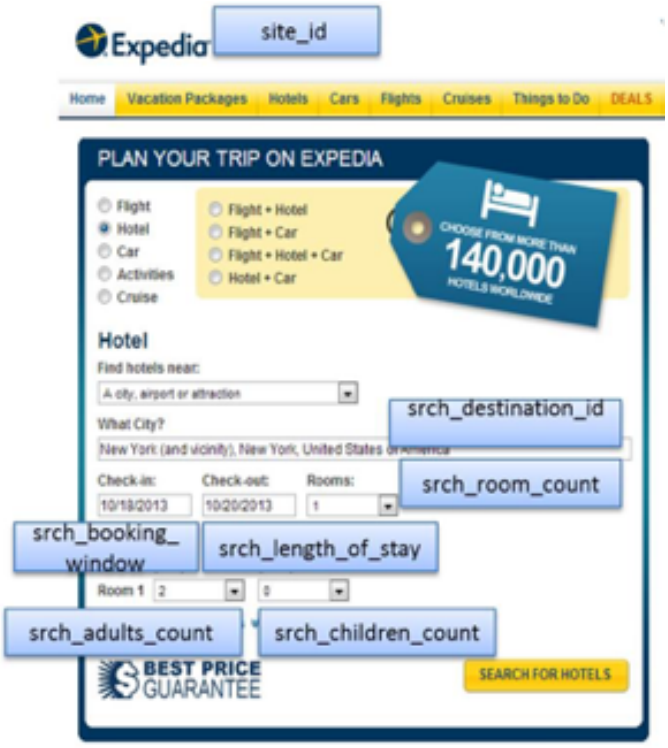

Figure 1: Search window

[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.

[3] Again based on Kaggle

Table 1: Description of the dataset (cf. Kaggle)

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
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 comp2_rate_percent_diff ... comp8_rate comp8_in comp8_rate_percent_diff		(same, for competitor 2 through 8)
		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

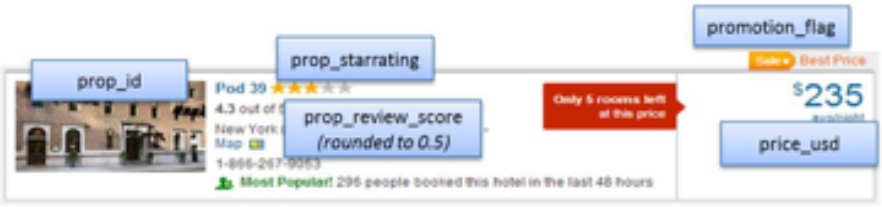

Figure 2: Hotel result

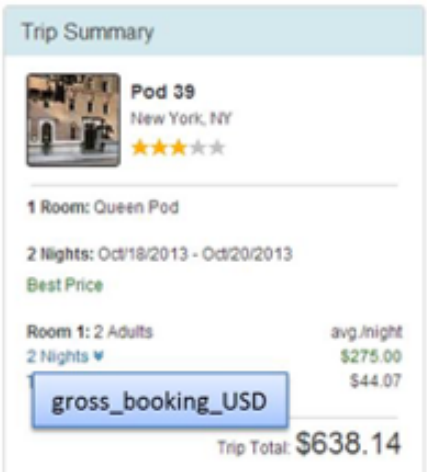

Figure 3: Cost overview

DETAILED TASK DESCRIPTION

To make things easier, we will use a DM process model to describe your task in a bit more detail, similar to what you have seen in assignment 1, and during the lectures.

TASK 1: BUSINESS UNDERSTANDING

Your task is to predict what hotel properties that result from a search of a user, the 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 in a 'Related Work' section in your report.

TASK 2: DATA UNDERSTANDING

In this task, you will do some exploratory data analysis (EDA). Explore the dataset, count, summarize, plot things, and report findings that are useful for creating predictions. Remember that EDA is not necessary done once at the start of the project. It is expected that you do some EDA, build some features, train some models, then some idea comes up, do some more EDA, modify your features, train another model on these new features, and so on.

TASK 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 the test set.
Of course, you are also allowed to use external data sources if you find ones that are useful. In case you get inspired by previous work on this competition, please make sure you properly cite the sources you base your approach on.

TASK 4: MODELING AND EVALUATION

Naturally, once you prepare the dataset, you should be able to build models. You have great freedom to select the techniques you feel are most appropriate and should try at least two different techniques. At least use one technique that has been discussed during the lecture on Recommender Systems or a variant thereof. . The choice of that techniques or alternatives you try might be influenced by how we would like to measure your predictions at the end (described later in this document). To test how your model is compared to other, you can upload your answers for the test set on the in class Kaggle website: `https://www.kaggle.com/competitions/dmt-2024-2nd-assignment`, see previous instructions for signing up using `https://tinyurl.com/join-dmt-cup`. Note that the score on Kaggle is only for part of the test set, the score on the rest of the test set will only be disclosed on the final lecture and will form part of your final grade.

TASK 5: DEPLOYMENT

While we do not go into real deployment, we do want you to go into a brief discussion on how Expedia could use your approach to deploy it on their systems in a scalable way, knowing that they have much more data available and also that the characteristics of the data can change over time. Discuss this in the light of the methods that have been introduced in the big data engineering and big data infrastructures lectures.

DELIVERABLES

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

PREDICTIONS

You'll need to submit your prediction file on the in class website of Kaggle, 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
```

Please make sure to submit as a team with team name in the format `VU-DM-2024-Group-x`, replacing `x` with the group number from Canvas, e.g. `VU-DM-2024-Group-132`. This way, we can take your score into account when grading and when identifying the winner of the competition.
The deadline for submitting the predictions on Kaggle is 19/05/2024, 23:59.

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. A discussion on scalable deployment of your approach.

3. 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.

4. Please format the document according to the LNCS guidelines. Templates are available on Canvas for both LateX and Microsoft Word, do not deviate from these templates. Note that you do not need to include an abstract in your report. The paper **should not exceed 14 including all figures and tables, but excluding references** (references do not count for the number of pages to encourage you to cite all relevant work). With the page limit, the aim is to challenge you to report only what is necessary. Make sure we can identify your report, i.e., your group number, names and student numbers should be in the document's header.

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 (using the same template) which addresses:

1. A schedule describing 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. Angie was responsible for transforming the dataset into a suitable format for the algorithms chosen whereas Berend was working on the report).

3. A critical reflection of the overall cooperation within the team.

Evaluation and Grading

Here's how you will get rewarded for your work. 80% of this mark can be achieved by submitting a nice and thorough report, 20% will come from where you end up in the competition. The process report will be used to make sure all contributed enough. In case of a clearly unequal contribution, grade differentiation will be applied within the group. The deadline for the reports is 24/05/2024, 23:59. The two reports should be submitted via Canvas while your prediction file should be uploaded on our Kaggle competition site with deadline 19/05/2024, 23:59. Regarding the competition-based marks, scores will be computed based on your results: the winner gets a 10, and a performance equal to random gives you a score of a 4. The evaluation of the competitions is explained in more detail below as well as the final presentation session. Furthermore, a detailed grading scheme can be found in Table 2.

Winning the competition

The winner will be rewarded with the fame and glory of winning the 2024VU 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 (NDCG)@5 calculated per query and averaged over all queries with the values weighted by the log_2 function. See `https://en.wikipedia.org/wiki/Discounted\_cumulative\_gain` 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. You should upload your prediction on the in class Kaggle website: `https://www.kaggle.com/competitions/dmt-2024-2nd-assignment`.

CLOSING EVENT

The presentation of your final assignment will be done during the closing event. 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. Six groups will be asked to present, those groups that ended up in the top 3 and three additional random groups will be asked to present their work. The closing event will take place on May 28th 2024 between 15:30 and 17:15 in room HG-01C (Aula).

Table 2: Grading scheme

Task	**Grading Component**	**Weight**
Scientific report	Dataset statistics	10
	Plots	10
	Rationale and interpretation	10
	Dataset pre-processing: report a replicable process of feature engineering	5
	Rationale for feature engineering	5
	Algorithm: which/why/how it works	10
	Parameters of algorithm used	5
	Evaluation of model created	10
	Quality of the writing	10
	Final model description	10
	Deployment in real life with big data engineering and infrastructure	10
	What you learned	5
	Extra page	-10
	Wrong formatting	-10
	Total	100 (Weight 80%)
Kaggle ranking	Total	100 (Weight 20%)
Total		100