ML Take Home Assignment

Welcome! We are glad you’ve decided to apply for a DS/ML role with us. In this exercise you’ll be able to showcase your skills tackling a search challenge. You can download the dataset here.

The dataset

In the [search.csv](https://drive.google.com/open?id=1DMkbZHfQ-31Futd0yxYqF-ATgB5i6IYY) file you’ll find over 175K entries of user search activity data. A brief description of the columns is presented at the end of this prompt.

The problem

Your manager believes that improvements to the search ranking algorithms are long overdue. The current solution bases ranking on most popular results for each query, but new sources of data open new opportunities to understand and customize results depending on user behavior.

Prior studies suggest session information is highly relevant, so assume all the user activity data is available through history – i.e., from the time of the query and before, not after – and any required transformation can be performed real-time in an efficient manner.

On the other hand, the product team has also heard about the project, and they’ve surfaced one big issue with the current strategy – new products are not shown within the best results! Given that fresh content is highly valued by customers, this is a big problem.

You have been tasked with -

1. Performing exploratory analyses to include new features and transformations
2. Redesigning the transformation pipeline for input data
3. Redesigning the search algorithm
4. Define a strategy that supports new content
5. Defining and justifying performance metrics

There’s no preference regarding architecture, so you can use either a top-k or a ranking approach, though you are expected to justify your choice.

The hand-in

Even though you are part of a technical team, the plan has to be approved by an executive panel, so a clear storyline, good visualizations and business acumen are a big plus!

We expect you to submit your work within a week, and though we use Python and Jupyter Notebooks ourselves, please feel free to use the language and format of your preference. Just remember to also include your code and comment it properly.

Data schema

|  |  |
| --- | --- |
| **feature** | **description** |
| id\_search | unique ID of the search |
| label | search result action (click, add\_to\_cart, purchase, impression) |
| id\_user | unique ID of the user |
| id\_product | unique ID of the product |
| ts\_search | timestamp of the search |
| ds\_search | date of the search |
| ds\_purchased | date when the product was purchased |
| ds\_cart | date when the product was first added to cart |
| query\_market | market of user search (e.g. Santiago) |
| query\_delivery\_date\_start | searched delivery start date |
| query\_delivery\_end\_date | searched delivery end date |
| query\_delivery\_radius | search query delivery radius |
| query\_price\_max | maximum price search filter |
| query\_price\_min | minimum price search filter |
| product\_is\_new | product is new (has 0 reviews and purchases) |
| product\_total\_price | total price of product for selected quantity plus delivery costs |
| product\_review\_rating | average review rating of listing given by guests (1 to 5) |
| product\_review\_count | number of guest reviews |
| product\_type | product type id |
| product\_category | product category id |
| product\_recommended\_items | number of related products recommended in product page |
| product\_has\_video | if product page has a professional video |
| product\_recent\_comments\_count | number of recent comments in product page |
| product\_recent\_comments\_count\_90d | number of comments in product page during the last 90 days |
| product\_volume\_discount | if listing has monthly discount |
| product\_bundle\_discount | if listing weekly discount provided by host |
| product\_short\_term\_price\_factor | product dynamic price multiplier (short term) |
| product\_long\_term\_price\_factor | product dynamic price multiplier (long term) |