# LDSSA Hackathon #5 - Recommender Systems

## Schedule

|  |  |
| --- | --- |
| ***Hour*** | ***Activity*** |
| 8:00 | Arrival, Breakfast, Student setup |
| 8:30 | Hackathon Prompt, Team Assignment |
| 9:00 | **Starting hacking!** |
| 12:00 | Goal - make first submission |
| 12:30 | Lunch Served (no need to stop hacking) |
| 14:00 | Goal - make improved submission |
| 16:00 | Start Working on Presentation |
| 17:00 | **Stop Hacking!** and Presentations Setup |
| 17:15 | Team Presentations |
| 18:20 | Instructor’s Presentation |
| 18:30 | Winners Announced & Beer time! |

## 

## 

## Overview

**In our last hackathon, you’ll build a recommender system for Sports products!**

## Objective

**The goal for this hackathon is to provide recommendations for a set of test users, based on historical reviews data.**

For each test user, you’ll provide a list of **50 recommended products**, sorted from the “most recommended” to the “less recommended” products.

Your submission file should be a csv with two columns:

* **user\_id**: the id of the test user
* **purchases**: list of recommended products

When you submit your predictions, some validations will be run that will check the following:

* Your file has the two columns with the right names
* Your file has the right number of rows
* Your file has the same user ids as the test dataset. Your submission will be sorted by user\_id on the portal side, so the order of the user\_ids on your submission file doesn’t matter

## Data sources

In the hackathon directory (in the students repo, you know the drill), you will find a data/ directory, with the following files:

* **train\_reviews.zip** - start by unzipping this file. It has reviews that users wrote about sports products. This dataset has the reviews both in text format and as numerical rating (in the **overall** column)
* **train\_products\_metadata.csv** - this file has some metadata about the products that were reviewed. In particular, you’ll find data about the product brand, categories and price
* **test\_users.csv** - this file has the list of users for which you’ll have to provide recommendations for
* **sample\_submission.csv** - Submission file example.

|  |
| --- |
| Note that, on your predictions file, each list on the **purchases** column should be converted to a string using the **json.dumps()** function. |

## Recommendations

❗Make sure that you get to and submit a baseline ASAP! Then work on improving it. In particular, it may take you some time to nail the correct format of the submission file, so don’t leave your submissions for the last minute!

❗There are test users for which we don’t have previous data. You’ll have to handle this, as you’ve learned on the BLUs.

❗You might notice that the reviews data has timestamps. These can be used when creating the validation dataset for local evaluation of your model (like we did on the binary classification hackathon and on the timeseries hackathon).

❗You’ve learned about multiple methods to provide recommendations, so it might be a good idea to divide and conquer, i.e, different team members can work on different recommendation strategies.

## Evaluation criteria for your model

Evaluation Metric - MAP@k, with k = 50.

You’ve learned about this metric in **BLU12 - Workflow**. If you feel that you need some clarification about this metric, [this article](https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52) may be useful, in particular the example on section 2.

**Disclaimer**

In the leaderboard, the scores you’ll see are the the MAP@k ‰ (i.e, MAP@k \* 1000). This is because we’re expecting the scores to be very small, and we just display 4 decimal cases in the leaderboard!

So, don’t worry if you see values higher than 1 for the score, because, in fact, the highest possible score is 1000.

## Hackathon Rules

* The selection of the teams is **random**.
* Instructors will be available to help at any time. The instructors will **not** help your team solve the challenge but they will help your team to be on track and answer technical questions that your team might have.
* **No more submissions and questions** to the instructors shall be done after the end of the challenge.
* Your team will have to prepare a presentation to share your findings with everyone. See the presentation guidelines [below](#_vkhpoxzb4mjg). This presentation will be considered in the overall evaluation of your team, so don’t consider it less important than the ML model!
* You can submit your predictions up to **five** times to evaluate your MAP@50 score. The best will be chosen for the team’s best score.
* The **final rank** is calculated as:

|  |
| --- |
| *FinalRank = 0.5 \* MAP@50\_rank + 0.5 \* Presentation\_rank* |

Where:

* ***MAP@50\_rank*** is the rank of your team in the leaderboard, considering the score of your **best submission**
* ***Presentation\_rank*** is the rank of your team in the presentation evaluation

The teams will be sorted by *FinalRank* ascending, and the first team wins!

Feel free to ask any questions about the scoring function!

## Presentation guidelines

* The presentation can take a **maximum of 4 minutes**. This is a hard limit! We’ll literally silence you and move on to the next group after the 4 minutes have passed.
* The presentation should approach the following topics:
  + Problem description
  + Data science workflow
    - Data preparation (EDA, what data did you use on your models, what preprocessing did you have to do before using the data)
    - Model selection (which models did you try, how did you evaluate them)
    - Results and discussion (which model did you end up choosing and how good was it)
  + Recommendations / Future work if you had more time to work on the problem
  + A funny pun at the end (not mandatory, but everyone loves it)!
* You can use [this template](https://docs.google.com/presentation/d/1g38u2K8WoYXEfHZNiFNmHCinUOUyd7gGSQBp54HLjug/edit?usp=sharing) if you want (make a copy of it and edit your copy).
* Charts/tables/great visuals are encouraged in your presentation. We actually have an evaluation criteria for the presentation which is “Used **relevant** visuals” (note the relevant!)
* The team can decide who is presenting. There are no rules here, you can go with one person presenting everything or having everyone presenting a part

## Procedure

* **Marking attendance:** go to the hackathon page on the [portal](https://portal.lisbondatascience.org/hackathons/student/hackathons/HCKT02/) and flag yourself as present in the hackathon. There is also an option if you’re joining remotely.
* **Taking presences:** instructors will double check that you are here, so expect to hear your name shouted by an instructor during this phase
* **Team selection:** you’ll be assigned to a team **randomly** by the portal
* In the hackathon page, you’ll have to select a **name** and **gif** for your team. This is what will be displayed on the [leaderboard](https://portal.lisbondatascience.org/hackathons/student/hackathons/HCKT05/leaderboard/).
* It’s also in this page that your team will **submit prediction files**.
* Send **@Hugo Ferreira** your presentation through Slack (shared link or pdf file) before “Submission of presentations”.

## After the hackathon

The leaderboard will be reopened after today so that you can keep trying to improve your score!

Also, next week we’ll share with you the code for a possible solution for the problem.

# Summary of Resources

* [Github repo](https://github.com/LDSSA/batch3-students/tree/hckt05-rec-sys/S05%20-%20Recommender%20Systems/HCKT05%20-%20Building%20a%20Sports%20Products%20Recommender%20System)
* [Leaderboard](https://portal.lisbondatascience.org/hackathons/student/hackathons/HCKT05/leaderboard/)
* [Template for Presentation](https://docs.google.com/presentation/d/1g38u2K8WoYXEfHZNiFNmHCinUOUyd7gGSQBp54HLjug/edit?usp=sharing)

Good luck!

*Lisbon Data Science Starters’ Academy team*

## 