CUHK-STAT3009 Project 2: Recommender systems based on real dataset with side information

Submission

- Prediction Submission in Kaggle
- Jupyter notebook submit to Blackbroad

Note: the submission in Kaggle must be consistent with the notebook submitted to Blackbroad, otherwise the final score will be zero.

Kaggle Submission

- Submit your final solution into Kaggle (https://www.kaggle.com/t/0d52d48031d341779acf3e4e4e92dd1d) (You must log in using this link).
- Use your student ID for all team members as your team name. For example, "1155111111; 1155111112; 1155111113".
- Don't cheat! Immediately fail the course if there is any cheating among groups.
- (**IMPORTANT!**) I revise Scored Private Submissions as 2, that is, you can include two submissions as your private leaderboard evaluation. Besides, if someone plays a trick like multiple submissions from multiple users, I will degrade your final score.
- (**IMPORTANT!**) Your final grade is based on the final private leaderboard at deadline, any action after ddl is **NOT** allowed! **Make sure you have changed the team name as your team members' students ID before ddl**.

Notebook Submission

The notebook should consist of following components:

Team members

At beginning of the notebook, you should include "Student Name" and "Student ID".

Contribution

The contribution of each team member should be clearly stated in the Notebook

Exploratory data analysis (`EDA`)

- Checking the description of the datatsets like the `data types` , how `many users` , `items` , etc
- Visualization on some important parts like `most rated items` , `most popular items` , `most rated users` , `frequency of ratings`
- Any missing data? any irregular data like `nan` , or `np.inf` ?
- How to pre-process the irregular data.
- Any other factors can help you make prediction

Model building

You may try many models and pick up the best one. I recommend you to introduce your final model as the structure as follows.

- Attempt model 1: (i) Which model you want to use; (ii) Any hyperparameters? how to tune; (iii) performance in Public Leaderboard; (iv) Any issue? (v) how to make improvement.

- Attempt model 2: (i) Which model you want to use; (ii) Any hyperparameters? how to tune; (iii) performance in Public Leaderboard; (iv) Any issue? (v) how to make improvement.

- Maybe more attempts ...

- You final model: (i) Which model you want to use; (ii) Any hyperparameters? how to tune; (iii) performance in Public Leaderboard; (iv) Explain why you think the model is the best.

Result

- Print the `user_id` , `item_id` , and `pred_rating` , for the T-th record in the `test.csv` , where T is the last four digits of your student Id. For example, if your `student Id = 1155111111` , please print the 1111-th record.

- Print the `top-5` preferred items based on your `predicted_rating` for the `user_id` in the above question.

Grading

- You will receive bonus point if you work solo to the final project.

- You final score depends on the score in **PRIVATE LEADBOARD** and the submitted **JUPYTER NOTEBOOK**.
    - **PRIVATE LEADBOARD** We will actually follow the Competition Medals Policy in Kaggle (https://www.kaggle.com/progression), that is:

      Gold Top 10% Silver Top 20% Bronze Top 40%

    - **JUPYTER NOTEBOOK** Grading for the Jupyter Notebook depends on the overall quality of your notebook: [EDA (20%) + Model building (55%) + Result (5%) + Overall philosophy (20%)].
- Your final score will depend on the performance in Kaggle (Kaggle_Medals) and the quality of your notebook (NB_score).

```
def score(Kaggle_medals, NB_score):
    if Kaggle_medal == 'Gold':
        score = 1.1 * NB_score
    elif Kaggle_medal == 'Sliver':
        score = 0.9 * NB_score
    elif Kaggle_medal == 'Bronze':
        score = 0.8 * NB_score
    else:
        score = 0.75 * NB_score
    return score
```

- Make sure your notebook is readable, and even ready to publish. You can check an illustrative notebook in Titanic Classification (https://www.kaggle.com/startupsci/titanic-data-science-solutions).