MIE1624H Introduction to Data Science and Analytics – Winter 2021
Course Project
Due Date: 11:59pm, April 4, 2021
Submit via Quercus

Background:

For this project, you are responsible to complete a Kaggle competition. Each team will build a recommender system model to make predictions related to Amazon Music Reviews. Your team will be required to complete a project report on your work in the form of a consulting report, as well as a final presentation.

Deadline:

Submission deadline: April 4, 2021. Presentations are on April 6, 6:00-9:00pm during lecture time via Bb Collaborate.

Submission:

You will need to submit a Jupyter notebook with your code, a report and your presentation slides through Quercus. Kaggle submissions will be made through the provided link. You will also be required to provide the code used.

Late Submission:

No late submissions will be accepted.

Dataset:

train.csv.zip 150,000 reviews to be used for training. It is not necessary to use all ratings for training, for example, if doing so proves too computationally intensive.

reviewerID The ID of the user. This is a hashed user identifier from Amazon.

itemID The ID of the item. This is a hashed product identifier from Amazon.

reviewText The text of the review.

summary A short summary of the review.

overall The star rating of the user's review from 1 to 5.

reviewHash Hash of the review (essentially a unique identifier for the review).

unixReviewTime Time of the review in seconds since 1970.

reviewTime Plain-text representation of the review time.

category Category labels of the product being reviewed.

test.csv.zip 20,000 reviews to be used for generating the final Kaggle submission. All fields are the same as in train.csv.zip with the exception of the **overall** rating removed.

rating_pairs.csv Pairs (reviewerIDs and itemIDs) on which you are to predict ratings.

baseline.py A simple baseline that computes a user average and global average on training data then uses this to make predictions on test data. This code is given to demonstrate how to properly format predictions for uploading to Kaggle. A submission made with this code corresponds to the 'naive baseline' submission on the leaderboard.

Please do not try to collect these reviews from Amazon, or to reverse-engineer the hashing function we used to anonymize the data. Doing so will not make it easier to successfully complete the project. **We will require a working code for all submissions to ensure no violation of the competition rules.**

Kaggle Competition:

Your team will build a recommender system model to predict ratings related to music reviews on Amazon. Specifically, given a (user, item) pair and associated review data, we want to predict the review's star rating as accurately as possible. The performance will be measured with MSE. Solutions will be graded on Kaggle, with the competition closing at 11:59pm EST, Sunday April 4 (note that the time reported on the competition webpage is in UTC!). The leaderboard will show your results on half of the test data, but your ultimate score will depend on your predictions across the whole dataset. You must include your Kaggle team ID (click "download raw data" at the public leaderboard) in your submitted report, and the name on your Kaggle team (show on the leaderboards) must match your name on your report. The link to the Kaggle competition will be posted on Quercus.

Marking:

The project is worth 20 points (12 points for your analysis and report and 8 points for your business-oriented presentation).

The presentation will be graded as follows (8 total marks):

- 2 marks for organization and delivery (e.g. clarity, enthusiasm, poise)
- 3 marks for content (e.g. proper visuals, high-level ideas, answering questions)
- 3 marks for the business pitch (e.g. recommendations, the solution to the problem)

The analysis and the report will be graded as follows (12 total marks):

- 3 marks for the analysis (e.g. cleaning the data, visualizations, applying algorithms)
- 3 marks for discussion and insight (e.g. how your analysis contributes to the problem, making a decision, storytelling)

- 6 marks for your Kaggle performance. Your Kaggle performance will be graded as follows:
  - Your ability to obtain a solution that outperforms the leaderboard baselines on the unseen portion of the test data (5 marks).
  - Obtaining full marks requires a solution that is substantially better than baseline performance.
  - Obtain a solution that outperforms the baselines on the seen portion of the test data (i.e., the leaderboard). This is a consolation prize in case you overfit to the leaderboard. (1 mark).
  - Students with submissions ranked in the top 3 teams will receive a 2 point bonus mark.

Note that we will be checking submissions for similar or copied code and to verify competition rules were followed.

Written Report:

You will also write a report about the approaches you took. Your report should be 12 pt font and be up to 4 pages excluding references. This report will be submitted on Quercus and is due by 11:59pm, Sunday April 4. Remember to include your Kaggle team ID in this report or we will not be able to grade your submission. Your report should cover the following sections:

1. Describe how you processed your data and what features you used. Your exploratory analysis here should motivate the model you use in the next section.
2. Describe your model. Explain and justify your decision to use the model you proposed. How will you optimize it? Did you run into any issues due to scalability, overfitting, etc.? What other models did you consider for comparison? What were your unsuccessful attempts along the way? What are the strengths and weaknesses of the different models being compared?
3. Describe your results and conclusions. How well does your model perform compared to alternatives, and what is the significance of the results? Which feature representations worked well and which do not? What is the interpretation of your model's parameters? Why did the proposed model succeed why others failed (or if it failed, why did it fail)?

Every group member gets the same mark for the project. It is your responsibility to determine how you split the work inside your group.

What to Submit via Quercus:

1. Your Jupyter notebook with appropriate documentation for every step as well as the relevant data files (in addition to data files provided by the competition). Make sure that your IPython notebook runs on Google Colab.

2. A 4-page consulting report in PDF format that summarizes your findings and results (all graphs should have axes appropriately labelled, all visual materials should be understandable and the graphics of sufficient quality to be easily readable.) This report should be business-oriented and cover your problem more extensively than your presentation.

3. Your business-oriented presentation slides in PowerPoint and PDF formats. (Each group will present their findings and results during a 5-minute presentation with 1-2 minutes for questions live on Bb Collaborate. Presentations will be timed and stopped after 5 minutes.)