STAT 4609: Group Project

Department of Statistics and Actuarial Science
University of Hong Kong

Second Semester, 2020-21

This group project aims to provide students with more practical experience of using big data analytics learned from the class on a real-life problem. You will learn how to formulate a problem and apply relevant big data analytics tools in practice. Each project should be done in a team of 4 students, though a team can have at most 5 people and at least 3 people. Each team should choose a team name and appoint a team leader, who is in tasked with submitting the group project assignment online. Before the final project due date, there will be intermediate assignments to ensure proper progress before the final submission.

For this group project you must select one of the following data sets:

1. The Netflix Prize dataset.
2. CIFAR-10 image data set
3. Fake News Challenge dataset. (evaluate your model on the `competition_test` data.)

You will build a supervised learning model on the training data and evaluate your model using the test set data. While you may use the same data preprocessing step from the training data on the test data, you are not permitted to edit the test set data in any way, doing so will constitute academic dishonesty. Each team's final result will be posted on Moodle at the end of the assignment. Your team will not receive extra points for having the best score, but you are permitted to take pride in the result.

For this group project, you must study and understand the dataset by exploring it. Pay attention to the quality of data, the meaningful features, data distribution, and the types of variable values. Perform and discuss any necessary data cleansing and transformation. Choose appropriate data analytics tools and develop the necessary model upon the dataset. You must justify your modeling choices and how your choices explain the outcomes of the project. You also must compare at least *three* different models that we have learned about in the course. If there are any hyperparameters in your model, you must discuss how you selected these hyperparameters.

Project Report On May 2, 23:59, your final report for the project will be due. You must submit a `.pdf` file and a Jupyter Notebook containing your analysis. Your written report has no page minimum but is a maximum of 15 pages single spaced and should include

- A cover page with
  - Title of the project (at most 15 words)

    - Name of the team
    - List of team leader and team members (with student UID)

- Objectives of the project (presenting the background of the project, the problem of the study, and project objectives).
- Data Sources: Description of data and data preprocessing.
- Modeling Choices: Describe what models you used and why.
- The results of your analysis on the test set data
- Conclusions and Future Plans (What could you do to improve your model in the future? What problems did you encounter trying to run your analysis and how did you fix them?)
- References (such as research articles, books, book chapters, websites, etc.). This is not included in the page limit.

Grading of project progress report will be based on problem formulation, data description and analysis, grounded justification for choice of models, interpretation of findings, originality and creativity, clear and organized writing.

The Jupyter notebook submission must neatly organized and clearly demonstrate each aspect of your model discussed in your written report. The Jupyter notebook must be able to run without errors when submitted for the final project. The model should be neatly organized into different functions, as you have done in the homework assignments. Students must write the model without the aid of any pre-written machine learning toolkit packages (ex: `scikit-learn, gensim,` etc.), though you are permitted to use numerical computation packages like `numpy`, `scipy`, `pandas`, `PyTorch`, etc.

Project Presentation In the week of May 3 to May 7, students must give a presentation on their project. Each team will spend 10 minutes for presentation using Zoom. Each team member should contribute in the discussion. The presentation should introduce what the project is about, how the your team chose the final model used in project, interpret the results of the analysis, and finally the conclusions and reflections from the project. Grading will be based on the content of presentation and oral presentation skills.

Plagarism Students are reminded to read the HKU policy on plagarism here:

`https://tl.hku.hk/plagiarism/`.

You are not permitted to copy any other group's work either in the written discussion or in the code. You are also not permitted to copy other peoples' implementation of code that you may find on online. Groups that are discovered to violate the plagiarism policy will receive a zero score for the final project grade.