Northeastern University

Project Guidelines

CS 6220 — Data Mining Techniques
Summer 2021

Course Project

Data mining is about the methods and systems for extracting knowledge or insights from data. To bring together and apply the various methods covered in this course, you will work on a project that invokes some or all of the data mining process to extract knowledge or insights on a topic of your choosing. You will acquire and analyze the data, formulate one or more questions of interest, apply relevant methods, evaluate the results, and communicate your findings.

Project Team

Students may work individually or in teams for the course project. Teams may have no more than 3 members, with equal contributions expected from each member. Larger teams are expected to do a project of proportionally larger scope and work.

Project Milestones

There are a few milestones for your final project. It is critical to note that no extensions will be given for any of the project due dates. Projects submitted after the final due date will not be graded. If you anticipate any issues, please contact instructor as soon as possible. Unless otherwise announced, all submission deadlines are at 6:00pm PT on the assigned due date.

Tuesday, July 20: Project proposals due
Tuesday, August 10: Project presentations
Friday, August 13: Final deliverables due

Deliverables

There are several deliverables for your project that will be graded individually to make up your final project score.

Proposal

You start your project by forming your groups and letting the instructor know what topic you are interested in exploring. The proposal should be an informal written document, roughly 1–2 pages in length, submitted to the instructor via Canvas. The instructor will review the proposal with your team to develop a finalized course of action. For details on proposal grading, please consult the rubric below.

Project Presentation

Each group will explain their project in a 10–15 minute presentation to the class. You may use visual sources (e.g., PPT/PPTX) to assist in your presentation. Your presentation should clearly convey the project ideas, data, methods, and results, including the question(s) being addressed, the motivation of the analyses employed, and relevant evaluations, contributions, and discussion questions. Please consult the rubric for details on presentation grading.

Project Paper

Each project will be concluded with a project paper. Your paper should summarize your steps in developing your solution, including how you collected the data, what data understanding and preprocessing you performed, the model method you used or implemented, the evaluation employed, and the insights obtained. The paper should be at least 3–4 pages in length (excluding tables and graphics) and submitted in Word or PDF format. Optionally, you may submit a draft paper via Canvas to receive feedback from the instructor prior to the final paper deadline. Please consult the rubric for details on paper grading.

Grading

The project grade has three components: the proposal, presentation, and paper. 60% of the project grade will be based on your project paper. 30% of the grade will be based on your project presentation. The remaining 10% of the grade will be based on your project proposal.

Project Proposal Grading:

Title of Project:	5%	What's the title of the project?
Project Plan:	30%	What do you plan to do? What steps of the data mining process are involved?
Data Sources:	30%	What data do you plan to use? Where will this data be sourced or collected? Has this data been used for similar purposes previously?
Proposed Evaluation:	20%	How do you plan to evaluate your proposed method? How will you determine whether the method is successful?
Writing Quality:	15%	Clarity of expression (5%), organization (5%), and grammar (5%).

Project Presentation Grading:

Introduction:	15%	Provide context. What questions are being addressed? Are your questions descriptive or predictive? Is the problem supervised or unsupervised? What is the impact of solving this problem?

Data and Experiments:	25%	What data did you use? How was the data collected? What data analyze did you perform to understand the data? Did you perform data preprocessing?
Solution/Method:	25%	What did you do? Why did you choose this method? What tools and techniques did you use?
Evaluation and Results:	20%	What evaluation did you do? Why is this evaluation appropriate? Do your conclusions match your results?
Presentation Quality:	15%	Clarity of speaking (5%), organization (5%), and visuals (5%).

Project Paper Grading:

Introduction:	15%	Provide context and motivation. What questions are being addressed? Why are these questions interesting or important?
Related Work:	10%	What other methods have addressed these or similar questions? How do these methods differ from your method?
Solution/Method:	20%	What did you do? What tools and techniques did you use? Was any innovation attempted? How does your solution compare to similar solutions, if any?
Data and Experiments:	20%	What data did you use? Are your experimental methods reliable? What preprocessing was done the data?
Evaluation and Results:	20%	Did you properly evaluate your experiments? Did you test for statistical significance? Do your conclusions match your results?
Writing Quality:	15%	Clarity of writing (5%), organization (5%), and grammar (5%).

Submission Instructions

To submit your project deliverables, create a folder for your project in your GitHub repo and place your files in this folder. The folder should include your project paper (Word or PDF format), presentation materials, and all code and output used to generate results. If you are working in a team, also provide a text file that briefly lists the contributions of all team members. Once your files are on GitHub, generate a link and submit it via Canvas to the instructor. *Note: Only one team member needs to do this per project.*

Project Ideas

You are free to suggest the themes that interest you for the project. In my experience, I have found that we have often identified interesting project topics based on discussions. To help generate some ideas, you can find examples of potential project topics below.

1. Compete in a Kaggle challenge
    - Kaggle website
2. Investigate algorithmic “fairness”
    - Paper: “Algorithmic Fairness”
3. Predict congressional party affiliations
    - Propublica data
4. Develop a strategy for countering class imbalance in data
    - Workshop on Learning from Imbalanced Data Sets
    - Special Issue on Learning from Imbalanced Datasets
5. Implement cost-sensitive learning methodologies
    - Paper: "The Foundations of Cost-Sensitive Learning"
    - Paper: "Active Cost-Sensitive Learning"
6. Implement active learning methodologies
    - Paper on active sampling for class probability estimation/ranking
7. Implement a scheme for combining ensembles of classifiers
    - Paper: "Learning Ensembles from Bites"
8. Topics in graph-based data mining or relationship learning
    - Paper: "Link Mining: A Survey"