CSC 3220 Team Project Fall 2023

DESCRIPTION

The team projects for this class are to be done in **groups of 3-4 students**. The idea is to perform a pseudo end-to-end data science project on a realistic task[1]. To this end, below is a list of recommended projects. Your group will specify your preferences among those projects. You can also propose a different project if you wish. Teams will be assigned to projects after reviewing each team's submitted prioritized list.

PROJECT SUGGESTIONS

NOTE: Specific projects you can **NOT** select, are listed at the end of this document in **Appendix A**!

- **Kaggle Competitions**: https://www.kaggle.com/competitions
- **JFK Assassination**: Records (https://www.archives.gov/research/jfk)
- **Driven Data for Social Good**: https://www.drivendata.org/competitions/
- **Data Hack:** https://datahack.analyticsvidhya.com/
- **Challenge Data**: https://challengedata.ens.fr/challenges (out of France but instructions in English)
- **UCI Repository** (https://archive.ics.uci.edu/ml/datasets/)

The sequence of stages in the project process is as follows:

PROJECT TEAMS AND PREFERENCES (due by **MIDNIGHT on Monday, October 30**)

Your designated team representative (henceforth termed "you" or "your") must submit (into the corresponding iLearn dropbox) your (1) project team list (groups should be not less than **3** students and no more than **4** students, and if not, *you must have a very compelling reason* for the exception), and (2) your complete ranked order of project preferences.

List your team's project preferences in order from first choice to last choice. Each item should include a number (1. = first preference, 2. = second preference, 3. = third preference, etc.) and then the name of the project. For projects from the above website, ALL YOU NEED TO SPECIFY in your ranked list is the name of the competition/dataset, and the URL to the competition/data science problem (because the website will define the challenge/problem).

If your team wants to do their own project, you must put "Own Project" in your ranked list, and then include the following additional information. There is no need to do the following IF your preferences are all from the above website list:

[1] Again, this is a pseudo-real-world project. In reality, it would take longer than the 5 weeks you have to do this assignment. You will not be expected to have a complete, detailed data science report – but hopefully you will gain some valuable experience to help you understand how team data science projects work.

Project Title

The name of your project.

Project Overview

Explain your project at a high level. What problem is it addressing? What question are you trying to answer? How will you measure success? Be as specific as possible, there should be some quantifiable measure that you can use to measure the success of your project compared to other systems or approaches.

Dataset

Describe the dataset(s) that you plan to use. How will you get it? How large should it be to answer your questions?

Tools

List any specialized tools that will be used (beside "R" – which is required for all projects), and whether you already have them or want to use some other resources.

PROJECT ASSIGNMENTS (announced on **Tuesday, October 31**)

In class, I will let every team know their assignments.

PROJECT: PROBLEM STATEMENT AND DATA (due by class time on **Thursday, November 9**)

Your **first deliverable** will be a complete draft of the following two sections (using the template found in **Appendix B**):

1. **Problem Statement and Background (15%)**
2. **Data and Exploratory Analysis (15%)**

Document must be submitted into the corresponding iLearn dropbox. ONLY ONE SUBMISSION PER TEAM INTO iLearn.

PROJECT: METHODS AND TOOLS (due by **Thursday, November 16**)

Your **second deliverable** will be (1) any updates/corrections to the previous sections, and (2) a complete draft of the following two sections (again, using the template found in **Appendix B**):

3. **Methods (10%)**
4. **Tools (10%)**

Document must be submitted into the corresponding iLearn dropbox. ONLY ONE SUBMISSION PER TEAM INTO iLearn.

PROJECT: RESULTS AND APPENDIX – FINAL REPORT (due in iLearn by class time on **Tuesday, November 28**)

Your **third (and final) deliverable** will be (1) any updates/corrections to the previous sections, and (2) a complete draft of the last three sections (again, using the template found in **Appendix B**):

5. **Results (35%)**
6. **Summary and Conclusions (10%)**
7. **Appendix (5%)**

The document must be submitted to the corresponding iLearn dropbox. ONLY ONE SUBMISSION PER TEAM INTO iLearn.

IN-CLASS PRESENTATION (due in iLearn by class time on **Tuesday, November 28**)

Use the following format, which should result in 6-8 slides plus a title slide.

0. Please identify your team (team name and member names) on a title slide.

1. (1-2 slides) Problem statement: What problem you were trying to solve? Should include quality metrics you used to measure performance/accuracy. Should *not* describe the algorithm or method you're using to solve the problem.

2. (2) Methods you explored. Include some data preparation/featurization, then the learning algorithms you tried, and possibly visualization or interaction methods.

3. (1 slide) The tools you used (in addition to R), and a rationale for their use. Can cover data preparation, learning, visualization, performance measurement(s), etc.

4. (1-2 slides) Results and unexpected challenges.

5. (1 slide) Summary and conclusions (including lessons learned).

PEER EVALUATION (due by **noon on Friday, December 1**)

Each team member will submit an evaluation of not only their teammates but themselves as well. This will be done through Qualtrics (*instructions to follow*).

A. APPENDIX - PROJECTS YOU CAN *NOT* USE

The following are projects that are **NOT** available to be selected:

- Kaggle:
    - “Titanic: Machine Learning from Disaster”

    - “Real or Not? NLP with Disaster Tweets”
    - “Video Game Sales”
    - “Digit Recognizer”
    - “Predict Future Sales”

    - “Anime Recommendations Database”
    - “House Prices”
    - “Air Pollution”
    - “Forest Cover Type Prediction”
    - “Predicting Blood Donations”
- Analytics Vidhya:
    - “Joke Ratings”
- DrivenData:
    - “DengAI: Predicting Disease Spread”
    - “Richter's Predictor: Modeling Earthquake Damage”
- Challenge Data:
    - “Stock Trading: Prediction of Auction Volumes”
- UCI Repository (https://archive.ics.uci.edu/ml/datasets/)
    - “Absenteeism at Work”

B. APPENDIX - PROJECT REPORT TEMPLATE

Your project report is the formal description of your project. Your report will be graded on: Quality, Completeness, Creativity, and Grammar.

The report should use the following template (**replacing the text in [brackets] with your text**.) The percentage worth of each section is shown in (parenthesis) at the end of each section header – do NOT include those percentages in your section headers.

[NAME OF PROJECT]

[TEAM MEMBER NAMES]

1. Problem Statement and Background (15%)

[Give a clear and complete statement of the problem. (Do NOT describe data, methods or tools yet – see below.) Where does the data come from, what are its characteristics? Include informal success measures (e.g. accuracy on cross-validated data, without specifying ROC or precision/recall, etc.) that you plan to use. Include background material as appropriate: who cares about this problem, what impact it has, what implications better solutions might have. Included a brief summary of any related work you know about.]

2. Data and Exploratory Analysis (15%)

[Describe the data set you will be using. Discuss anything you had to do clean the data and why. Describe what tools and R code you used to extract, clean, and generate the data for your experiments. Some potential questions of the data might be: any anomalies or outliers? Did you need to impute any of the data in order to get it to work for any proposed algorithms?]

3. Methods (10%)

*[Describe the methods you explored (usually algorithms, or data cleaning or data wrangling approaches). Justify your methods in terms of the problem statement. What did you consider but *not* use? In particular, be sure to include every method you tried, even if it didn't "work". When describing methods that didn't work, make clear how they failed and any evaluation metrics you used to decide so.]*

4. Tools (10%)

[Describe the tools that you used and the reasons for their choice. Justify them in terms of the problem itself and the methods you want to use. Tools will probably include machine learning, and possibly data wrangling and visualization. Please discuss all of them. How did you employ them? What features worked well and what didn't? What could be improved? Describe any tools that you tried and ended up not using. What was the problem?]

5. Results (35%)

[Give a detailed summary of the results of your work. Here is where you specify the exact performance measures you used. Usually there will be some kind of accuracy or quality measure. There may also be a performance (runtime or throughput) measure. Please use visualizations whenever possible. Include links to interactive visualizations if you built them. You should attempt to evaluate a primary model and in addition a "baseline" model. The baseline is typically the simplest model that's applicable to that data problem, e.g. Naive Bayes for classification, or K-means on raw feature data for clustering. If there isn't a plausible automatic baseline model, you can e.g. compare with human performance by having someone hand-solve your problem on a small subset of data. You won't expect to achieve this level of performance, but it establishes a scale by which to measure your project's performance. Compare the performance of your baseline model and primary model and explain the differences.]

6. Summary and Conclusions (10%)

[In this section give a high-level summary of your results. If the reader only reads one section of the report, this one should be it, and it should be self-contained. You can refer back to the "Results" section for elaborations. This section should be less than a page. In particular, emphasize any results that were surprising.]

7. Appendix (5%)

Include the link to your github/gitlab repository (that I can access) containing your R programs/scripts, and link to the data.