BME 590.01 Course Project — Intro to Biomedical Data Science

Deliverable Due Dates:

Groups Finalized.....11:59 PM ET, February 11th, 2021
Project Proposal ..... 11:59 PM ET, March 4th, 2021
Midterm Presentations..... 8:30-9:45am ET by Zoom, March 18th, 2021
Project Final Report ..... 5:00 PM ET, April 20th , 2021
Project Presentations ..... 8:30-9:45am ET by Zoom, April 20th & 22nd, 2021

Overview:

The best way to learn about using data science is to actually *use it*. The purpose of this course project is to give you an opportunity to explore elements of the data science pipeline from start to finish, from preprocessing, data exploration and visualization, predictive model design and selection, and effective communication of results. Even more importantly, this is an opportunity to explore the techniques on an application that's interesting to you.

The course project will be completed in groups of 2 or 3. Your course project should address an interesting and meaningful problem. Inspiration can come from your future career goals, a personal hobby project, ongoing research, etc. The project's topic is fairly open as long as it pertains to Biomedical Data Science in some way, given that (i) the problem is difficult enough to be interesting (i.e. linear regression doesn't give you 100% accuracy) and (ii) the problem is feasible (i.e. the data exists and there are relevant methods to address your question). **The goal of the initial project proposal** is to check that the proposed idea fits these two criteria. Be creative about the problem you want to address; the project should be something that engages you!

Each group is expected to:

1. Form project teams. Please discuss with members of the class before or after class times (you can hang out in the open area on Gather.town to meet your classmates) to identify your project partners. Please meet with your team at least two times before February 11th, 2021. One member per team should upload a pdf document to Sakai by 11:59pm ET on February 11th, 2021 listing: 1) the group members first and last names, and 2) the date and time of your meetings, and 3) a brief 1 paragraph summary of the project ideas that were discussed at your initial meetings. This will count toward 20% of your proposal grade.
2. Identify a data source to use: You can use any data you like that is at least loosely related to biomedicine. There are lots of publicly available data sets online. Potential data resources are listed in our class slides from Unit 1. You can also refer to the UCI machine learning repository (http://archive.ics.uci.edu/ml/index.php) or the Kaggle website which lists datasets that have been used for machine learning competitions.

You can also use a dataset that one of you are using for your research - if you go this route make sure you have the correct IRB permissions and permission from the PI of the research study to use the data for this class.

- A good dataset is one that is large enough to find some interesting patterns but not so large and messy that it takes you forever to work through. UCI is useful as most of those datasets are slightly curated and have descriptions as well as associated papers.

- Some datasets may require you to apply for access. Please begin this process *immediately* as this can take some time. By the time of your proposal, you should have access to the data you want to use.

3. Three primary tasks after the data set is chosen:

1. Perform exploratory data analysis - this will involve describing the predictor variables (e.g. distributions, missingness) and looking at any underlying structure (e.g. PCA, Clustering)
2. Develop a prediction model - implementing different prediction models and apply the validation scheme
3. Evaluate the model - identify important variables and evaluate performance and visualize results and communicate your findings and interpretation

I do not expect you to invent new algorithms as part of this project. Your goal is to show that can apply techniques from class and explore additional material related to your question/application.

Note that the project will be evaluated on the approach; not on the metrics of the results. In the real world, conclusions from data science techniques are often **negative**. There is information in a negative result, and you **will not** be penalized for negative results, as long as the problem was approached in a reasonable way and the methods were approached and evaluated correctly.

Details on Deliverables:

There are 4 different deliverables associated with this project. As mentioned above, group formation is designed to occur early to give you and your groupmates time to brainstorm and identify data sources. The project proposal is designed to check that the proposed project is appropriate. The proposal, including the group formation, is 5% of the total grade. The purpose of the proposal is to make sure that groups are 1) formed and 2) are on track to succeed. The final report is 20% of the total grade, and is expected to be a complete and comprehensive report. Each group will give two presentations for the midterm and final, which will account for 10% and %15 of the final grade, respectively. Each of these (pdfs of documents or slides) will be submitted electronically as a pdf (including your presentation) to Sakai by 1 group member, following the file naming convention of the homeworks:

date_lastname_firstname_[group/proposal /finalReport/presentation].pdf (**5 points will be deducted for files named incorrectly**).

Project Expectations:

To give a sense of the expectation of the final project, the final report is expected to be complete reflection of the work on the project, and should be modeled on something you could hand to a manager or research advisor. *If you are graduating soon, think about this as an opportunity to have a portfolio-quality item that you could show to a potential employer.* Page limits will be given on individual assignments, and the report should be concise and self-contained with enough information to recreate and reproduce what you have done.

The content of the final report will be highly individualized to your project, but it should contain material to fully address the following categories:

- Problem Description: Background and context for the problem you have selected, and the motivation for considering it for this project. Some questions here to think about are: why is this an interesting problem? Why might others be interested in this problem? What is the *exact* question that you're trying to answer? What metrics will be used to assess your performance on this question?
- Data description: Provide a description of the data you used, where it came from, and why it will help you answer the question you're asking from the data. What are considerations (e.g. sources of error) for the data type you chose to work with? Note any peculiarities of the data and how it would influence the analysis procedure that you're choosing.
- Data preprocessing: What preprocessing approaches were chosen, and why? How did those choices relate to the type of data chosen? What were their effect on the metrics of performance? What is the sensitivity of the final result to the choices made here?
- Chosen algorithmic pipeline: What approach did you choose? Why did you choose this approach? How sensitive is the final result to the chosen approach?
- Performance results: What are the estimates of the performance metrics from your complete pipeline on the dataset? How does this help answer the question you're trying to ask from the data? Provide curves that address trade-offs in model selection.
- Conclusions: What is your overall assessment of the system you built? What are its strengths? What are its weaknesses? What might you have done differently if you had an opportunity to redo the project, and why? What answer to the question you asked does the data support?