PES University, Bengaluru
UE18CS312 - Data Analytics

Session: Aug – Dec 2020
Project FAQ's

1. **How large should the dataset be (in terms of samples and attributes)?**

   Ans: There is no restriction per se, but we definitely want you to have some 'insight' to the problem and solution approach. Unlike a typical solution approach in machine learning (or deep learning) or perhaps even big data, we do not expect you to work with voluminous data that would be thrown into a black box to get some stellar output. We want you to have an understanding of the data you are working with and be able to interpret the results and analyze them in some detail.

2. **Does the final accuracy not matter?**

   Ans: While we encourage you to build meaningful models that are indeed as accurate as possible, we do not attach as much importance to the final outcome as we do to the solution approach. We want you to explain the rationale behind your solution approach and analyze the results – what are the assumptions made prior to the design and to what extent does the data conform to those assumptions? What do the results mean? Why did model X not work for data Y? What could have been done to improve upon the model?

3. **Must we necessarily work with some data set that is publicly available or are we allowed to collect our own data or work on a problem that is not necessarily data intensive?**

   Ans: You are most welcome to collect data or work on a problem that does not require EDA in the traditional sense (analyzing customer feedback or product, movie or book reviews to arrive at a numeric rating, etc.). However, pleasure ensure that all your effort will not go into acquiring the data or learning NLP, image/ video, etc., processing, since the class project has to be about analytics!

4. **Can we use Machine learning models?**

Ans: Yes you may. As long as all the effort is not in getting a 'black box' (various deep neural network architectures) to work, but you do have some time to interpret the results and analyze them and get into the merits and limitations of the solution approach, this will not be an issue.

5. **My data set has only two features; is it ok?**

Ans: If you are working on stock price prediction, for example, there are plenty of problems you can investigate (such as trends, seasonality in the data, etc.); so yes, it is ok in principle. However, in general, if you have a number of features for a classification (recommendation) type problem, you will have more room to understand the interaction of features (study scatterplots, look at correlations, etc.).

6. **We are working on a recommendation system. What should we keep in mind before we begin?**

   Ans: Who are you building the recommendation for? What sort of recommendations would you like to make? How will you evaluate the strength of the recommendation? These are some of the points you could keep in mind. You could study predecessor approaches paying special attention to these points.

7. **When we look for data or work on a problem statement, what should we bear in mind?**

   Ans: How old/ new is this data? If the data is old, what has already been tried and tested on this data? How is what we propose going to be different from (or improve upon) what others have already done? There is a certain problem statement associated with the data (13 features for prediction of housing prices, for example). Is there any other creative question we can ask on this data that the data was not originally collected for, but has enough information within it to answer? How will we evaluate our solution approach (or test our models)? Is any visualization going to help bring out patterns in the data or throw insight to the interaction between variables or help us visualize the results better (like a time lapse of change in housing prices or a heat map of traffic blocks at various times in the day), etc.?

8. **What should we look for in papers for literature review?**

Ans: You can look for papers on the (a) the same data set (b) the same problem domain but different data sets (c) similar problem domain (movie recommendation given the preferences vs book recommendation given the preferences) or (d) papers pertaining to a similar solution approach (use of regression models for forecasting in general, may be not necessarily for the same dataset) or working on a recommendation system with high class imbalance, etc.

9. **Can we use R or Python for the project?**

Ans: Yes, you are welcome to use either R or Python (or, if absolutely necessary, a combination of both). If you are keen to explore anything else, do let us know. (We need to have the expertise to evaluate your work to permit you to go ahead with a programing language/ visualization tool!)

10. **We were late in submitting the project teams. Will we lose marks?**

Ans: Timeliness of submission and consistency of work are two key components of the evaluation. No marks has been deducted for this first submission of team details. Hereafter, please do ensure you plan for the submission of the milestone/ deliverable or status check and submit to us whatever the progress is – does not matter if it not yet as ‘perfect’ as you would like it to be – the fact you have thought about it and are working on it will matter.

Deadlines

11. **What are the intermediate submissions (tentative) we need to be prepared for?**
Ans: Here is the list

September 16: Working title and coarse problem statement (what is the problem statement? Why is this useful/ how is this different from what exists? What is the data set you will be working on/ collecting?)
September 23: A brief summary of three papers (what is the paper about? What are the assumptions in the paper? What is the solution approach? How is this relevant to your work?)
September 30: A brief summary of the EDA effort on the data and/ or a summary of three more relevant papers and any refinement to the assumptions on the data, solution or evaluation strategy and division of tasks – who will be responsible for what? (for the upcoming weeks)
October 12: Team and individual Git accounts + Literature review report (details of format and content in the project guidelines document) + team weekly update – Blog/ Slack/… (optional)
October 21: List the preliminary models designed/ implemented (who implemented what?) what are the results? What next?
October 28: Progress with newer solution approach (who tried what?) What is the evaluation/ validation strategy used/ that will be used?
November 11/16: Progress with newer solution approach (who tried what?) What is the evaluation/ validation strategy used/ that will be used? How good are the results (or not)?
November 18/23: Checklist of submission elements: **Draft of the report, code+data, video script should be ready for submission –** refinement alone remains to be done (adding comments to the code + any additional parameter tuning that may help, beautifying an interface, if any, adding sections on discussion of results in the report (corresponding to parameter tuning or any newer models to be tested), etc.)

*****