CS 5644: Project Description

One of the main goals of CS5644 is to prepare you to apply the analytics skills learnt in this course to 'real-world' datasets. Use the steps below to structure your project planning and design.

Step 1:

Students should form teams of 1-3 people and work on their project. (Since the lower bound is 1, this means you can choose to do a project by yourself. But be forewarned: you will have to do all the work on your own if you choose this route and cannot plead for leniency in evaluation). The first task is to form a team. Use the Student Discussion forum (read especially the Student Introductions discussion) to form such teams.

Step 2:

Once a team is formed, pick a project topic or subject domain. E.g., the subject domain can be insurance, medical records, political science, etc. Then, what is the problem that you are attempting to solve in this domain? Who cares about the problem? Why is this problem best approached as a data analytics task? For instance, if the data comes from a mortgage provider, perhaps the analytics question to predict which individuals are likely to (or not) default on their mortgage. If the data comes from a hospital, perhaps the analytics question is to predict which patients are at risk of getting readmitted. If the data comes from a large retailer, maybe the analytics question is to predict the monthly sales of some product(s). These are just examples and there are numerous possibilities.

Step 3:

Convince yourselves (and us!) that there is a real-world dataset in your domain available to be mined. (Thus, go back to Step 2 if a suitable dataset is not available). The dataset can be an open (i.e., publicly available for free) dataset, for example: (not all of these fit the size requirements but are given as an example).

- The New York Taxi dataset,
- The Enron email dataset,
- Twitter network dataset,
- Facebook dataset,
- Movies and sentiments database, and so on.

Some good dataset repositories:

- UCI Machine Learning Repository
- Kaggle Datasets
- Awesome Datasets Repo

There are zillions of such free datasets out there. Alternatively, if you are already working with a dataset (obtained via your employer or otherwise) on which you could conduct analytics, you are welcome to do so. Please ensure that you are not committing any violations pertaining to information ownership, privacy, or copyright. Finally, please carefully read the evaluation criteria below to determine if a dataset is large

enough. Sometimes you can create more features through feature engineering (i.e. bag of words modeling or one hot encoding of categorical data columns).

Step 4:

What exactly is the machine learning problem for this dataset in your domain? For instance, this could be a supervised machine learning problem like: classification, regression, or it could be an unsupervised problem, e.g., clustering, association analysis.

Deliverables and deadlines:

- **Step 1 : Proposal** Form a team, decide the topic you are going to work on, and the details above. Talk to the teaching staff if necessary. The team as a whole submits a 1-2 page proposal (11 point font, 1 inch margin) fleshing out the project. Who are the team members? What is the domain? What are the data science insights you are seeking? What dataset are you hoping to study? Do you have access to it? Identify some simple characteristics of your dataset, e.g., for instance: how many data points are in the dataset? What are the features? How many features? What type of features are they? (Numerical? Categorical? Boolean?) What is the target you are trying to predict (in case of a supervised setting)? Which algorithms/techniques/models you plan to use? What do you expect to submit/accomplish by the end of the project? i.e., What will be improved using the results of your analysis? How will you measure/evaluate such improvement? (For instance, if you are trying to predict customer retention using machine learning, how is customer retention currently predicted? You will need to know the current state-of-the-art in order to know whether your machine learning approach is performing better/worse.)

- **Step 2 : Project Milestone** Think of this stage as your final project report but without your major findings. Expand your 1-2 page proposal above into a 2-3 page report. You should have teased out the data at this point. You should have identified and extracted features for your experiments. You ideally should have tried out some machine learning algorithms on a small subset of your dataset.

- **Step 3 : Final Project report** This is the final report with all your findings, results, etc. This should typically span 6-8 pages. It should contain extensive experimentation, and the details of the analysis and insights gained from the analysis should be presented. For instance, if you are solving a classification problem, you should try multiple algorithms, e.g., decision trees, nearest neighbors, NBC, etc. You should compare and contrast the results and explain the insights learnt through these algorithms.

Evaluation criteria:

- **Impact and design:** You should pick a 'real' problem. What impact would the work have? Is a data-driven approach the best way to tackle the problem? Who benefits with the data-driven solution? How well is the problem formulated/developed? Is it cast as a supervised learning problem or unsupervised learning problem? Why?
- **Scale:** The scope of the project should be beyond that of a homework assignment. You should consider a sizable dataset (e.g., **a few hundred thousand instances** with about **200-250 features**, like the examples above) and 'play' with the data – data cleaning, preparation, feature engineering, etc. The data would have to be 'massaged' before it is ready for a machine learning algorithm/setting.

- **Algorithms:** Did you try all possible algorithms/techniques to solve the machine learning problem that was formulated? For instance, if it is a classification – you must compare decision trees, logistic regression, support vector machine classifier, etc.
- **Insights/Inference:** What are the actionable insights? How do the features extracted from the data impact the observed solution?

Note:

- In a team project, individual contributions should be clearly specified.
- Kindly follow submission guidelines, failing which penalty might apply.