CSC-272: Big Data **Spring 2020**

DATA MINING NEWS JOURNAL

It is a primary objective of this class to relate the technical details of data mining and machine learning algorithms to real world applications involving human behavior. As we have read in Data Mining so far, such algorithms can have a profound impact in terms of not only *predicting* what humans will do, but gaining a deeper understanding of their behavior. Another objective is to encourage you as lifelong scholars to maintain a habit of regularly consulting authoritative periodicals to stay abreast of what's happening in your field in society at large. This ongoing assignment is designed to contribute to these objectives.

Starting today, there is a standing weekly assignment to find one article or current event concerning data mining/predictive analytics. The key criteria for the stories that you select are *currency*, *authority*, and *relevance*. That is to say, they should describe current research or application of data mining, should come from a respected, vetted source, and be accessible and of interest to students of data mining. All selections should be related to some aspect of human behavior. They should also be about *hidden knowledge* rather than s*hallow knowledge*. Please keep in mind that just because a study uses data, that doesn't mean it is data *mining*.

Sources for this assignment abound. Try to use different sources throughout the semester. You may consult newspapers such as the *New York Times*, *Washington Post*, and even the local paper if you find something interesting. You may consult periodicals such as *Time*, *Newsweek*, *Science*, and *Scientific American*. It is not required that the source be a peer-reviewed academic journal. The Furman library has subscriptions to nearly every periodical you can imagine, and search facilities that should make finding a story easy.

There are numerous web sites devoted to data mining and the broader topic of data journalism. Get in the habit of regularly consulting Nate Silver's FiveThirtyEight data journalism site (http://fivethirtyeight.com/) as well as KDnuggets (http://www.kdnuggets.com/) for news stories. Also visit the Kaggle competitions page (http://www.kaggle.com/competitions) for examples of data mining problems being solved currently. Find other similar sites of your own.

Though blogs generally are anathema to academics, for this assignment they can be a valid source – as long as you establish the credibility of the author. All three sites mentioned above have an associated blog (http://blog.kaggle.com/ for example, which re-caps the results of recent competitions). A simple Google search for "data mining blogs" will provide a rich source of possible stories.

If you enjoy listening to podcasts, this is another avenue you may explore for this assignment. FiveThirtyEight publishes two podcasts – "Politics", featuring stories about how data is affecting our political behavior, and "The Lab", which applies analytics to stories in the world of sports. (I listen to both of these each week and have learned a ton from them.) You may use episodes of these or any other data analytics-themed podcasts to satisfy this assignment.

Given this wealth of sources, you should be able to avoid repeating the same source two consecutive weeks. If for some reason you feel you must, be prepared to argue why that was your only option.

A new entry to your journal will be due at 5pm every Thursday evening, unless I explicitly cancel the assignment for a given week (which I will do, for example, when an exam is coming up). Entries are to be posted to the CSC-272 Facebook group, and are to include (1) a **link to your selected story**, (2) a **100-150 word summary/explanation** of the story, and (3) a **100-150 word**

reflection (guidelines below). Entries may not be repeated, so it is important that you both read the submissions of others, and that you not procrastinate.

Please put a paragraph break between the summary/explanation and the reflection.

Each reflection should have the following qualities:

- It should be in your own words. That is, you should select a story that engages and interests you, so that you are able to describe its contents without using phrases or complex terminology copied from the story itself.
- It should go beyond the "what?" question and also discuss the "how?" As in: "How is the data mining accomplished?" in your story. What does the dataset look like (attributes, instances, class attribute)? What methodology (data collection, data cleaning, machine learning technique, evaluation, etc.) is used? Relate the story directly to things we are learning in class.

Reflections should also try to answer one of more of these questions:

- How does the story affirm or amplify something we've learned in class? Or refute something?
- What impact is the application having, or might it have, on society? What does it stand to teach – or predict – about human behavior?
- Are there any interesting innovations to describe, or possibly even flaws that you find questionable?
- In what other problem domains might the application be fruitfully applied?
- What makes it interesting to you?

Your entries will not be graded on grammar or elegance of phrasing. This is a participation/effort assignment for the most part. If you understand the objectives outlined above, you will find it fairly straightforward to meet those objectives.

In class Friday, the discussion will be organized around key themes that emerge from the collective group of posts. A component of the grade for this assignment will be your participation in this discussion. If you report a particularly interesting story, or report it particularly well, I may ask you to give a short presentation in class. Similary, your story may be used to help make a larger point in support of a theme that emerges.