Sana Khan

Homework 1

Part 1

1. Discuss whether or not each of the following activities is a data mining task.
   1. Dividing the customers of a company according to their gender.
      1. No, this is an example of querying a database
   2. Dividing the customers of a company according to their profitability.
      1. No, this is also an example of querying the database
   3. Computing the total sales of a company.
      1. No, This is also an example of querying a database
   4. Sorting a student database based on student identification numbers.
      1. No, this is another example of querying a database
   5. Predicting the outcomes of tossing a (fair) pair of dice.
      1. Yes, since we are predicting the outcome
   6. Predicting the future stock price of a company using historical records.
      1. Yes, since we are predicting the outcome
   7. Monitoring the heart rate of a patient for abnormalities
      1. Yes, because we would have to have an understanding of what normal is based on previous data. We would then use that to detect what is abnormal
   8. Monitoring seismic waves for earthquake activities.
      1. Yes, because you would have to have an understanding of what is normal based on previous data. This could then be used to detect abnormalities
   9. Extracting the frequencies of a sound wave
      1. No, because it isn’t being used to predict new information
2. Describe how classification, clustering, association rule mining and anomaly detection can be used at a company
   1. Classification can be utilized by an internet search company to help provide more results that are similar to what the user is searching for. For example if a user searches SUV, through classification we could have results of mid-size SUVs also populated based on the category
   2. Clustering – this groups similar items into a bucket that would then be presented to the user based on their search terms.
   3. Association Rule mining – uses patterns to determine that two variables may be related. This could be helpful in cross sell or upselling a user. For example if a user searches ‘washing machines’ then detergents may also be displayed
   4. Anomaly Detection - is used to determine if a variable is very different than other data. An example would be if an otherwise uncommon term has become very popular. This could be used to then advertise new information to the user.
3. For each of the following data sets, explain whether or not data privacy is an important issue.
4. Census data collected from 1900-1950.
   1. Yes, census data is private data the contains confidential information about the users who filled the survey.
5. IP addresses and visit times of Web users who visit your Website.
   1. Yes, this data could be used in a malicious way if the wrong person had access to it
6. Images from Earth-orbiting satellites.
   1. No, this does not contain confidential information unless it is very detailed
7. Names and addresses of people from the telephone book
   1. No, this is publicly available information that does not contain confidential information about the users
8. Names and email addresses collected from the Web.
   1. No, this is publicly available information that does not contain confidential information about the users

Part 2

Top of Form

Bottom of Form

# The article *Google Flu Trends: The Limits of Big Data* written by Steve Lohr for the New York Times is a criticism of using big data analytics to predict flu trends. The article describes how for more than one flu season Google Flu Trends has overshot the prediction of the number of positive cases. This article references *The Parable of Google Flu: Traps in Big Data Analysis* which claims that Google believes “big data sets trump traditional data collection and analysis”. These authors believe that CDC data, which may have a time lag, is more accurate and reliable than Google Flu. The authors have suggested a combination of using Goolge Flu and CDC data to have better results in predicting flu seasons, something Matt Mohebbi one of the inventors of Google Flu Trends also agrees with. Mohebbi in the article claims that this tool was intended to be used alongside other data to predict trends.

The article In Defense of Google Flu Trends written by Alexis C. Madrgial is a defense of Google Flu Trends and how it uses big data to work. This article explains that the intended use of Google Flu Trends is to complement CDC data plus other data sources to try to predict flu cases in the season. The article also explains how some of the suggestions of combing Google Flu Trends with CDC data is not ideal, because this would take away from being able to use the tool to understand certain trends. GFT can be used to understand when “data reflects a real deviation from the standard trend, or whether it is an artifact of the model.” This article also mentions how Johns Hopkins used GFT along with other data sources and it was the only model that provided “statistically significant forecast improvements over the base model.”

I believe that the article written in favor of Google Flu Trends provides an example of not believing the hype around big data. It urges readers to trust the CDC, which has tried and true data collection/analysis methods than GFT. I do believe there is validity to this because as end users, we should look into different data sources or different perspectives. I do feel that the NY Times article is taking GFT a bit out of context. The second article focuses on explaining how GFT should be used alongside other data to complement or improve the data. Some benefits of having a tool like this are easy access to vital information during an epidemic, or a pandemic like Covid. A potential problem is if the tool isn’t vetted it could lead to providing inaccurate information that people could make decisions off of. I think that large companies can do a better job of relaying the risks to the end user.