Assignment 01

Task 1: Review Data Mining Concepts and Tasks

1. Discuss if whether or not each of the following activities is a data mining task:
    a. Dividing the customers of a company according to their gender.
        i. **This is not a data mining task. This is a trivial task and the data in this case (the customers) are already known.**
    b. Dividing the customers of a company according to their profitability.
        i. **This is not a data mining task. Once again, the data in this case (the customers) are already known and you are performing a simple calculation based on an accounting metric.**
    c. Computing the total sales of a company.
        i. **This is not a data mining task. The total sales of a company can be determined by a simple calculation. This is too trivial to be a data mining task.**
    d. Sorting a student database based on student identification numbers.
        i. **This is not a data mining task. There is no actual insight or information being extracted through this activity.**
    e. Predicting the outcomes of a tossing (fair) pair of dice.
        i. **This is not a data mining task since it is a (fair) dice toss and we are not estimating the outcomes of random tosses.**
    f. Predicting the future stock price of a company using historical records.
        i. **This is a data mining task being that we could create test and train data to predict the future stock prices. This is predictive analytics and a time series analysis would work as a good technique to model the future stock prices.**
    g. Monitoring the heart rate of a patient for abnormalities.
        i. **This is a data mining task. As we learned in this lesson, anomaly detection detects significant deviations from normal behavior. We can build a model for this which monitors the normal heart rate and alerts us whenever there is an unusual/abnormal heart rate.**
    h. Monitoring seismic waves for earthquake activities.
        i. **This is a data mining task using anomaly detection we can build a model to monitor the normal seismic waves for earthquake activities and set an alert for abnormal seismic waves based on a wave amount.**
    i. Extracting the frequencies of a sound wave.
        i. **This is not a data mining task being that we already know the sound wave frequency.**

2. Suppose that you are employed as a data mining consultant for an internet search engine company. Describe how Data Mining can help the company by giving specific examples of how techniques such as clustering, classification, association rule mining, and anomaly detection can be applied.
    a. As a data mining consultant for a **digital marketing firm,** I would us the following techniques:
        i. Clustering: We can cluster customers into four (4) or more segments based on what products they search on a client's website.
        ii. Classification: We can assign each new customer to a cluster based on the probability that they will fall under that particular cluster based on past and current product searches.
        iii. Association Rule Mining: A customer's cache can be set to identify which queries are associated with other queries when customers visit their site.
        iv. Anomaly Detection: A client's website can be analyzed to detect which pages are most popular and which are less visited. We can then modify and/or delete the pages which seem to be least visited by potential customers.
3. For each of the following data sets, explain whether or not data privacy is an important issue.
    a. Census data collected from 1900-1950.
        i. **No**. Data privacy is not an important issue because all information collected through this medium wouldn't be used in sampling of the data. Additionally, they are general data collected from all persons.
    b. IP addresses and visit times of Web users who visit your website.
        i. **Yes**. This is a data privacy issue because the personal data of visitors to the site is being collected. Their data must be safeguarded.
    c. Images from Earth-orbiting satellites.
        i. **No**. Since this is public to all users, this data is not private.
    d. Names and addresses of people from the telephone book.
        i. **No**. This is public to all users and is not private.
    e. Names and email addresses collected from the Web.
        i. **No**. This is made public to all users of the internet and thus is not private.

Task 2: Practice Your Critical Thinking and Writing

Criticism

Google Flu Trends (GFT) had many expectations and was believed to set the precedent for what big data analysis is capable of. Unfortunately, critics claim that those expectations were not met. This particular article focuses on the article, "The Parable of Google Flu: Traps in Big Data Analysis", written by authors Alessandro Vespignani, a professor at Northeastern University, David Lazer, a professor at Northeastern, Ryan Kennedy, an assistant professor at the University of Houston, and Gary Kin, Harvard University's Institute for Quantitative Social Science Director. Upon further analysis of the experiment, it was revealed that GFT's service overestimated the amount of reported flu cases in the 2012-2013 flu season and in the last few years. Additionally, GFT's "estimate for the 2011-12 flu season was more than 50 percent higher than the cases reported by the Centers for Disease Control and Prevention…and for a period of more than two years ending in September 2013, the Google estimates were high in 100 out of 108 weeks." (Lohr, 2014) GFT is accused of having "big data hubris", defined as "the implicit assumption that big data sets trump traditional data collection and analysis." (Lohr, 2014) Even after modification, the service improved but still overestimated by approximately 30 percent. Lastly, GFT is accused of not using a broad array of data analyses tools.

Defense

The 2008 *Google Flu Trends* experiment was, in its own way a success. The experiment aimed to "predict the prevalence of the flu from searches that users made for about 40 flu-related queries." (Madrigal, 2014) It was praised on the news media from outlets such as The New York Times, The Wall Street Journal, and CNN. Following the notoriety in the news, Google Flu Trends was met with criticism from different outlets claiming the initiative had not lived up to the goals set forth at its inception. More specifically, GFT was believed to have claimed the capability of predicting future CDC reports. While ambitious, Google and CDC authors specifically stated that, "the system is not designed to be a replacement for traditional surveillance networks or supplant the need for laboratory-based diagnoses or surveillance." (Medrigal, 2014) Additionally, "as with other syndromic surveillance systems, the data are most useful as a means to spur further investigation and collection of direct measures of disease activity." (Medrigal, 2014) A team at John Hopkins built a practical influenza forecast model based on real-time, geographically focused and accessible data. At the conclusion of this research, the team concluded that "Flu Trend data was the only source of external information to provide statistically significant forecast improvements over the base model." (Medrigal, 2014) Since its launch,

researchers and scientists have been able to use Flu Trends model as a means to reach their institutional goals.

Thoughts

While the criticism of GFT could have been a valid one, GFT was only guilty in the court of public opinion and expectations. It had accomplished what it set out to do, which was to create a complementary signal which syncs with other signals. The goal was never replacement, but a separate entity which could understand various epidemiological scenarios and detect flu trend anomalies when monitored; as shown in the John Hopkins experiment. Lastly, GFT contributed to the broader perspective of Big Data Analytics as a complimentary tool that could be used in traditional surveillance systems.

Works Cited

Lohr, S. (2014, March 28). *Google Flu Trends: The Limits of Big Data*. Retrieved from nytimes.com: https://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/

Madrigal, A. C. (2014, March 27). *In Defense of Google Flu Trends*. Retrieved from theatlantic.com: https://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/