2020-0701 IST 707 Data Analytics

Homework #1

Sathish Kumar Rajendiran

2020-07-07

Table of Contents

[Task 1: Review data mining concepts and tasks 2](#_Toc45221196)

[Task 2: Practice critical thinking and writing 4](#_Toc45221197)

## **Task 1: Review data mining concepts and tasks**

Answer the exercise questions 1-3 in Textbook 1.7. For Question 2, feel free to change the question scenario from “an Internet search engine company” to any organization that you would like to think of. It can be a company, government office, NGO, etc.

1. Discuss whether or not each of the following activities is a data mining task:

|  |  |  |
| --- | --- | --- |
|  | Activity | Is data mining task? |
| a) | Dividing the customers of a company according to their gender | Yes, Clustering |
| b) | Dividing the customers of a company according to their profitability | Yes, Clustering |
| c) | Computing the total sales of a company | No, Information retrieval |
| d) | Sorting a student database based on student identification numbers | No, Information retrieval/database querying |
| e) | Predicting the outcomes of tossing a (fair) pair of dice | No, Since the die is fair- it’s more of probability calculation |
| f) | Predicting the future stock price of a company using historical records. | Yes, Predictive modeling. Finding target variable of regression model |
| g) | Monitoring the heart rate of a patient for abnormalities | Yes, anomalies detection |
| h) | Monitoring seismic waves for earthquake activities | Yes, anomalies detection |
| i) | Extracting the frequencies of a sound wave | No, Information retrieval |

1. Suppose that you are employed as a data mining consultant for [xyz]. 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.

Data mining for Anaplan (SaaS based Planning Software Company) – Data mining techniques can help expand customer base for Anaplan with its subscription-based renewal process.

Storage:

Big Data system on the Cloud (AWS/GCP/Azure) can be used as a data lake.

Data:

Lot of raw data can be collected for the past 24 to 36 months from various systems about their customers, partners and prospects. Observations including Account type (customer, prospect, partner), Name, domain, website, physical address, Age of Customer/Account creation date (existing customer), Line of Business, Use Cases, Number of employees, Total Revenues, Company index (G2K, Public/Private).

External data like hiring trend from LinkedIn, Average tenure of employees, competitor details, stock exchange values etc.

Internal data (first-party data) including Opportunity type, Opportunity stage, Opportunity value and internal data (first-party data) including product utilization, customer engagement, Number of licenses sold, licenses activated, active licenses, Usage logs and customer support data including number of cases, issue areas, resolution times etc.

Predictive modeling:

Based on the variety of independent variables, Predictive modeling technique “classification” can be applied to predict target variable - whether the customer would churn or expand.

Association Analysis:

Find patterns that describe strongly associated features in the data. Such as company domain and use cases bought (Finance and HR, Sales Planning and Sales Finance etc.). Include recommendations based on the association analysis to help expand new use cases. example for manufacturing domain, recommend supply chain along with finance and accounting use cases.

Clustering:

Find groups of closely related observations. Such as group companies based on their firmographic, technography, geographic locations, LOB, Size, Revenues etc.

Anomaly detection:

Identify outliers earlier in the process whose characteristics are significantly different from rest of the data. Often Customer engagement score or number of cases escalated may turn out to be an outlier on few observations.

1. For each of the following datasets, explain whether or not data privacy is an important issue:

|  |  |  |
| --- | --- | --- |
|  | Activity | Is data privacy a concern? |
| a) | Census data collected from 1900-1950 | Public data |
| b) | IP addresses and visit times of Web users who visit your Website | Yes, it has potential PII information of the user. |
| c) | Images from Earth-orbiting satellites | Public data |
| d) | Names and addresses of people from the telephone book | Yes, it has PII information of the user. |
| e) | Names and email addresses collected from the web | Yes, it has PII information of the user. |

## **Task 2: Practice critical thinking and writing**

Read the following two news articles. One criticized Google Flu Trend, and the other defended it. Write one paragraph to summarize the criticism, and another paragraph for the defense. Write the third paragraph to offer your own thought, e.g. is the criticism valid? Does the defense make sense? What other problems or benefit do you see in Google Flu Trend or similar big data applications?

http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/

**Criticism:**

In 2008, Google released a forecasting model called Google Flu Trends (GFT) leveraging the power of Big Data, which attempted to predict the prevalence of the flu from searches that users made for about 45 flu-related queries. However, it received a lot of criticism mainly due to the hype it created and considered a failure model in public opinion. All 3 versions of the GFT models released subsequently in 2009, 2013 and 2014, were inaccurate and overestimated the number of flu cases in almost 100 out of 108 weeks and often estimating more than 50% higher than the cases reported by Centers for Disease Control and Prevention (CDC) confirming how poor the accuracy was. In addition, CDC reports from Doctors on influenza-like illness which lag by two weeks was a more accurate predictor than Google Flu trends.

http://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/

**In Defense:**

In 2008, based on data up to that point in time Google Flu Trends (GFT) worked really well in predicting flu trends from 45-flu related queries that users made. CDC which has been involved in the shaping of how it functioned, liked the data it produced as well. There were lot of criticism followed in the subsequent years as the model fell short of accuracy. However, the main objective of the project owners was to build a model that produces complimentary signal to other signals. In addition, it was observed that when combining GFT model with CDC’s lagged data resulted better value than either could provide alone. It was also proved later, when a team examined to build a better influenza model having GFT as the only source of external information with statistically significant forecast improvements over the base model. Thus, providing a complimentary signal and confirming its methods to be useful and relevant.

**Personal View:**

Google Flu Trends idea is a prime example of how a hype around technology innovation can both fail yet become a revolutionary in its field. Criticism around GFT was valid as the numbers fell short in great margin and it recommends the need for better data quality, feature engineering, overfitting, anomalies, noise reduction as other critical aspects of the release. However, tracking 45-flu related queries over billions of searches adding collective intelligence of society in real time free from human bias, traditional methods are proof that GFT as a pioneer in big data. Thus helping, researchers in laying a foundation on big data analytics.