Correlation One: Data Scientist Challenge

Business Context

Your client is an Internet financial media company that sends out a personalized daily digest email of links to users, selling ads against this content. The digest email process works as follows:

- Employees of the company identify **articles**, often by user submission. Each article has a **topic** (e.g. Trading, Taxation, Industry Reports), and a **type** (e.g. Blog Post, White Paper, Tweet, Comment).
- The company collects these links and periodically sends a **personalized digest email** to users with links to a set of articles. These digest emails are typically sent a few times a week to each user, with delivery times typically staggered throughout the day.
- Users can click through the links they receive to **view** the article.

The Business Problem

The company cares a great deal about the amount of interaction it has with the recipients of the daily digest email and they want as many users clicking through the content as possible. The ultimate health of the business depends on user engagement with high-quality content.

The company makes frequent changes to how they determine the right content to send to users, and they have prioritized many kinds of content from different sources over time. The company suspects that different kinds of article topics and types have different levels of engagement for different users, but have generally adjusted that mix of content through a combination of intuition and high-level metrics.

Now, the company wants to get serious about data analytics. **Your task is to determine if, and how, the likelihood of a user viewing a link they receive has changed over the three months of data the company has provided**.

Answer Submission Guidelines

Your response should be emailed to sham@correlation-one.com.

You should send three items, namely your

- Code used to answer the question. We should be able to run the code; you can assume the dataset is placed in the directory "./data/"
- Resume
- Answers to the questions below. You are free to write as little or as much as you like in response to questions—the goal is to supply all relevant information as concisely as possible.

(1) What is your one-sentence executive summary?
(2) What is your detailed assessment (for a technical audience)? Please quantify, use technical jargon.
(3) What tools did you use?
(4) What techniques did you try?
(5) What three plots did you make to best explain the data?
(6) What is your commercial recommendation for business unit heads who are non-technical?
(7) What other data would you like to see about the platform? What questions would this additional data help you answer?

Dataset Schema & Definitions

The dataset consists of two files.

The first is a gzipped SQLite database and can be downloaded here. The database schema is as follows:

- **users contains the user ID and email address for all users, both users who have contributed articles (authors) and users who have not**
- **topics contains the names of all article topics**
- **types** contains the names of all article types
- **articles has a row for each article submitted by a user**
  - **author_id references the user id of the user who submitted the article**
  - **topic_id and type_id reference the topic and type of article**
  - **submission_time is the time (in EST) when the article was submitted**
- **email_content has a row for each article link sent to a user**
  - **email_id is the id of the digest email (which contains multiple article links)**
  - **user_id references the user who receives the email**
  - **article_id references an article that the digest email contains a link to**
  - **send_time is the time (in EST) when the digest email is sent to the user**

A second dataset contains clicks on article links by users receiving digest emails. This dataset is a gzipped apache access log file and can be downloaded here. This file contains user clickstreams: for each click on an article link by a user, there is a row in the log file. The format of the file is

[%t] "GET /click?article_id=%d1&user_id=%d2 HTTP/1.0" %b %d3

where

- %t is time (in EST)
- %d1 is the article_id (which references article_id in the **articles table above)**
- **%d2 is the user_id (which references user_id in the users table above)**
- **%b is the status code that the server sends the client (see apache log file referenced above for more details)**

- %d3 is the byte size of the content returned by server (see apache log file referenced above for more details)