Data 1050, Fall 2022
Project

You work for a data analytics company, Brown Analytics, which has secured a new project to analyze advertising data. Your client is an ad exchange platform, Ads Are Us, which serves a number of advertising clients. You are the lead on this project. Ads Are Us places ads for a variety of consumer products on a number of online publishing platforms. They would like to have insights into the likelihood of an ad for a certain type of product being clicked on given the sentiment of the surrounding textual content and the gender and age group of the viewer. For instance, what is the likelihood of an ad for a dieting product to be clicked on if the surrounding textual content is positive and if the viewer is a female.

Based on the data provided to you, you need to come up with a set of recommendations regarding ads for which products should be placed in which kind of textual content taking into account the sentiment of the textual content, the gender and age group of the viewer of the ad. You are NOT required to build a predictive model or a recommendation engine. All you need to do is compute some correlations and make recommendations based on the observed correlations in the data.

Data provided to you:

You are provided with a log file (not a database) and a database containing two tables, which are described below.

The log file is in .csv format and consists of rows which contain the following type of information:

Sentiment, Publication_URL, Product URL, clickORnot, gender, age_group

(An actual log file will contain more information such as a timestamp, ip address, etc. But we don’t want to clutter it with information which we don’t need for our analysis.)

For example, a row may contain the following information:

Positive, https://nytimes.com, https://vitamix.com/blenders, 1, female, middle-age

Sentiment records the sentiment of the textual context in which an ad was displayed to a viewer. It is NOT the sentiment of the ad. Sentiment has 3 values: Positive, Neutral, Negative.

Publication_URL is the URL of the (online) publication in which an ad for a specific product (e.g., Vitamix blender) appeared. For simplicity, let us assume that even if the product has several models (versions) the ad is model-agnostic.

Product_URL is the URL of the brand’s page for that product on its eCommerce site. (e.g., https://Vitamix.com/blenders). It is intended to be the same URL as the Product_URL in the database table ‘products’.

clickORnot records whether the viewer to whom the ad was displayed clicked on the ad or not. You may assume it is 0 or 1, where 1 records that it was clicked on.

gender is the gender of the person to whom the ad was displayed (believe it or not, advertising platforms have this kind of information). Possible values for gender are male, female, non-binary.

age_group records the age_group of the person viewing the ad. The values are among ‘juvenile’, ‘young’, ‘middle-age’, ‘senior’. This can be encoded as 1, 2, 3, 4 (1 being juvenile and so on).

The database consists of two tables, *products* and *product_categories*

Attributes of products: Product-name, Product URL, product_type

Product-name records the product name, which includes the brand (e.g., Vitamix blender).

Product_URL records the page of the product on the brand’s page for that product on its eCommerce site. (e.g., https://Vitamix.com/blenders). For simplicity, all the different models of blender by Vitamix share this page. This is the gold standard in terms of which Product_URL entries in the log file are validated as correct or corrupted.

Product_type: The type of the product (e.g. blender, in our running example)

Attributes of product_categories: product_type, category

product_type: For example, blender or lipstick or computer (does not include the brand or model number). Same as ‘product_type’ in ‘products’ table.

category: the generic category of the product (e.g., small kitchen appliances or cosmetics or electronics)

You are required to do the following tasks:

1. Some of the Product_URLs in the log file might have been corrupted. Write a Python (or PySpark) procedure to determine which Product_URLs are corrupted. Let us assume that if a Product_url in the log file doesn’t occur in the *products* table, it is regarded as corrupted. Using this procedure identify and list the corrupted URLs. (10)

2. For each corrupted URL what will you do with it? Don’t assume that for each corrupted URL the correct approach is to delete that log entry. What if the URL contained ‘.cam’ instead of ‘.com’ but otherwise corresponded with a URL in the ‘products’ table? In that case the proper approach would be to correct the URL. In other cases, the URL might be so corrupted that the best approach would be to delete that log entry (the entire row). Describe your approach to dealing with corrupted URLs. That is, describe your approach to determining that a URL is too corrupted to be rescued. It must describe a) a procedure for determining the degree to which the URL is corrupted, b) a threshold for determining in terms of this degree of corruption whether it can be corrected, and c) for those which can be corrected, identifying its corrected form. For extra credit implement this in a Python (or PySpark) program. (25 + 20 points for extra-credit)

3. For each product, compute all the Publication_URLs containing an ad for that product. (Don't just give the results. Show all the work by which you got those results. This applies to all the questions below.) (10)

4. For each *product type*, compute all the Publication_URLs containing an ad for that product type. Your solution must be scalable. That is, it should work well even if there are hundreds of products in each product_type and there are hundreds of product_types. (Hint: To make it scalable you should consider using a Python or PySpark script instead of a SQL query.) (20)

5. Save this information in the database. Should you save it in the *products* table or the *product_categories* table or should you create a new table, *product_type_pubURLs*, and save this information in this table? If you create a new table, make sure to set up all the appropriate foreign key constraints. On the other hand, if you use one of the existing tables, explain how you will avoid redundancy in your data. In either case, justify your decision. (10)

6. For each product, compute the click rate for it. (Click rate is the number of times a display of an ad was clicked on (by any user) divided by the number of times it was displayed (to any user). That is, the click rate is not specific to each user.) (10)

7. For each product, compute the click rate for each sentiment type. (10)

8. For each product *type,* compute the click rate for it. (10)

9. For each product *type* compute the click rate for each sentiment type. (10)

10. Save this information you computed in 9 above in a database table. Should you save it in the *products* table or the *product_categories* table or the *product_type_pubURLs* table, or should you create a new table *product_type_sentiment_clickrate*, and save this information in this table? If you create a new table, make sure to set up all the appropriate foreign key constraints. On the other hand, if you use one of the existing tables, explain how you will avoid redundancy in your data. In either case, justify your decision. (10)

11. Determine if the gender of the person viewing ads make a difference with regard to the click rate of ads shown in different sentiment context. That is, determine if there are any 'significant' differences in the correlation between the sentiment type of the ad context and clicking on the product type conditioned on gender. You can decide if any difference counts as 'significant'. (This is *not* a yes or no question. Compute the different correlations.) (10)

12. The same question as 9 above but replace gender with age-group. (10)

13. Based on your results make your recommendations. These should be in the form:
    a. Based on our analysis (give details of your analysis), ads for such and such product are most likely to produce clicks in such and sentiment context (or state that we see no correlation between click rate of an ad for a product and the sentiment context of the ad)
    b. Based on our analysis (with details), ads for such and such product are most likely to produce clicks in such and sentiment context by viewers of such and such gender (or state that we see no correlation between click rate of an ad for a product and the sentiment context of the ad and the gender of the viewer).
    c. Based on our analysis (with details), ads for such and such product are most likely to produce clicks in such and sentiment context by viewers of such and such age-group (or state that we see no correlation between click rate of an ad for a product and the sentiment context of the ad and the age-group of the viewer). (15)

Deliverables:

A video presentation presenting your main recommendations to Ads Are Us and the methodology you used for your analysis. Pretend that this video presentation is for the benefit of a client (not your instructor). It should be no more than 10 minutes. (20)

A written report containing the answers to all the questions above. List any assumptions you have made. Explain what you have done to make your methodology scalable. State the limitations of your analysis, if any. (10)

Total points (not including the extra-credit points): 200

The data (log file and the database file) will be provided to you next Monday. Meanwhile, you can start working on this project using the demo data provided (in the Files/project folder on Canvas). Even though the data provided to you will be largish, the idea is to think in terms of a much larger data set. The methods you use should work well even for a very large data set.