DataHack 2022 Challenge

The Challenge

Your DataHack 2022 team is part of a data science group at a private equity firm. Leadership is interested in making a significant investment in Upworthy, a viral news publishing & aggregation website. However, in the past, Upworthy has been disparaged for relying on clickbait to generate clicks, exposure, and signups. In response to this, Upworthy pledged to stop creating clickbait headlines and articles. Your firm's executives want to invest in Upworthy, but they are skeptical about whether Upworthy genuinely changed their publishing practices or not. Using a rich dataset of A/B tests from Upworthy, your team seeks to answer the following questions:

1. **Did Upworthy really change their clickbait publishing practices?**
2. **What was the impact of that change?**
3. **Do you recommend an investment in Upworthy?**

Your job is to answer the questions and justify your answer using data. This problem can be tackled in a variety of ways, such as:

- Using natural language processing (NLP) to measure/quantify clickbait
- Using time-series tools to analyze website characteristics before/after Upworthy's no-clickbait pledge
- Analyzing impact in terms of financial/growth/ethical implications
- Building a compelling dashboard/visualization in the context of the key question

We highly encourage creativity! Feel free to tackle the challenge in any way as long as it answers the key questions.

Competition Deliverables & Evaluation

By the end of the competition, your team is expected to deliver **(1)** a concise 4.5-minute (or less) PowerPoint presentation about your idea and key findings, and **(2)** any code that you write/use. Judges will evaluate your work via 3 categories:
1. **Creativity:** How novel are your ideas? Is your submission exciting?
2. **Technical Mastery:** Is your proposal technically correct? Did you use any interesting technologies?
3. **Quality:** How effective is your presentation? How polished is your code? How well can you sell your idea?

Dataset Description

For your analysis, you are given 3 .csv files (**upworthy-archive**, **daily-user-info**, **country-data**). upworthy-archive is the main datafile which reports the A/B tests and results; daily-user-info reports pageviews, new users, and other statistics over the course of the data; country-data reports some aggregate statistics by country. The A/B tests in upworthy-archive are used to determine which combination of **article headline** + **image** maximize clickthrough rate on an article. For these tests, randomization occurs at the user-level, and the user is assigned to one of several potential **packages** (which is a bundle of article headline + image). See Figure 1 for details. For each package, click/impression statistics are reported. Then, based on the results of the test, an Upworthy editor manually selects which package to show (not necessarily the best-performing one) when the test is complete.

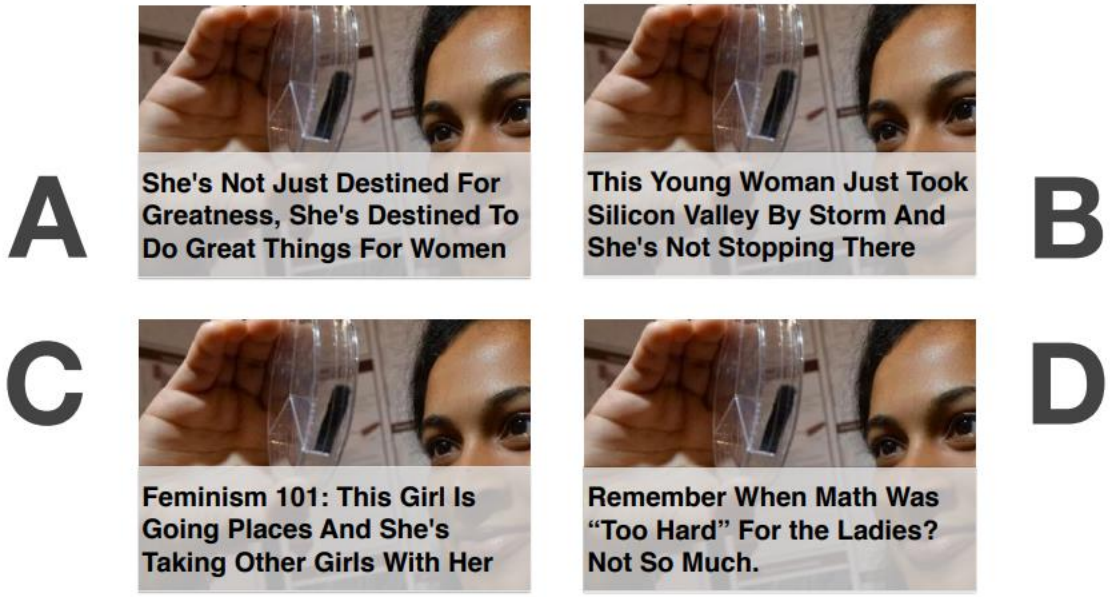

Figure 1: Users are randomly shown package A/B/C/D (or more variants). For each package, only the package_headline and package_picture_id varies. Packages in the same A/B test are not guaranteed to have the same image.

upworthy-archive.csv	
Feature	**Description**
created_at	When the package was created.
updated_at	When the package was last updated in the system.
ab_test_id	Unique identifier for an A/B test. Users were assigned to packages with the same ab_test_id.
package_headline	The article headline based on the package.
package_picture_id	Unique identifier for the picture based on the package. The picture itself is not included.

clicks	Number of users who clicked the article within an A/B test
impressions	Number of users who were exposed to a package within an A/B test
score	A score given to a package within an A/B test. We were unable to figure out how this score was calculated.
first_place	Shown to editors to guide which package to choose as winner
winner	Editors choose which package to show when the A/B test ends; not necessarily the best-performing one
excerpt	Excerpt from the article *
lede	Opening sentence of the story *
slug	Internal name for the web address *
share_text	Text shown when the article is shared off-website *
share_img	Image shown when the article is shared off-website *
test_week	Week the test was conducted

* clarifies the column does not vary by package; it is article-specific but not package-specific.

daily-user-info.csv	
Feature	**Description**
day	All statistics below aggregated by day
users	Number of unique user visits
new_users	Number of unique new user visits
sessions_per_user	
sessions	
avg_session_length	
bounce_rate	Percent of users who only viewed 1 page
pageviews	
pages_per_session	

country-data.csv	
Feature	**Description**
country	All statistics below aggregated by country
users	Total number of users
sessions	
avg_session_length	
bounce_rate	% of users who only viewed 1 page
pages_per_session	

What is an A/B test?

Knowledge of A/B tests is essential to succeeding in this competition (and as a data scientist). We recommend this resource for understanding what an A/B test is.

Some Tips

- Don't jump straight into the key question. Take time to deeply understand Upworthy and clickbait. Feature engineering is crucial.
- Incorporate information about the results of the A/B tests into your analysis—don't ignore this valuable information!
- For this competition, we recommend using DeepNote for effective collaboration and easy submission.
- Feel free to reach out to a mentor if anything is unclear—don't waste valuable time if you don't know something.

Online Resources Policy

Online resources are allowed, but you **must cite your sources** in the competition submission form (pasting the link is fine). Not doing so is **plagiarism** and will result in disqualification**.** Judges are familiar with what has/hasn't been posted online already, and they will be checking submissions to ensure compliance.