

Outbrain Click Prediction

## Can you predict which recommended content each user will click?

Yogita Jain | Sweta Bajaj | Antariksh Sharma

Advances in Data Science INFO 1790

Project Proposal

# **End User Persona**

**Outbrain** is going to be the end user here.

Currently, Outbrain pairs relevant content with curious readers in about 250 billion personalized recommendations every month across many thousands of sites. In this competition, Kagglers are challenged to predict which pieces of content its global base of users are likely to click on. Improving Outbrain’s recommendation algorithm will mean more users uncover stories that satisfy their individual tastes.



Sites

Content

Users

Aim: Improve outbrain’s recommendation algorithm which will uncover stories that satisfy user tastes.

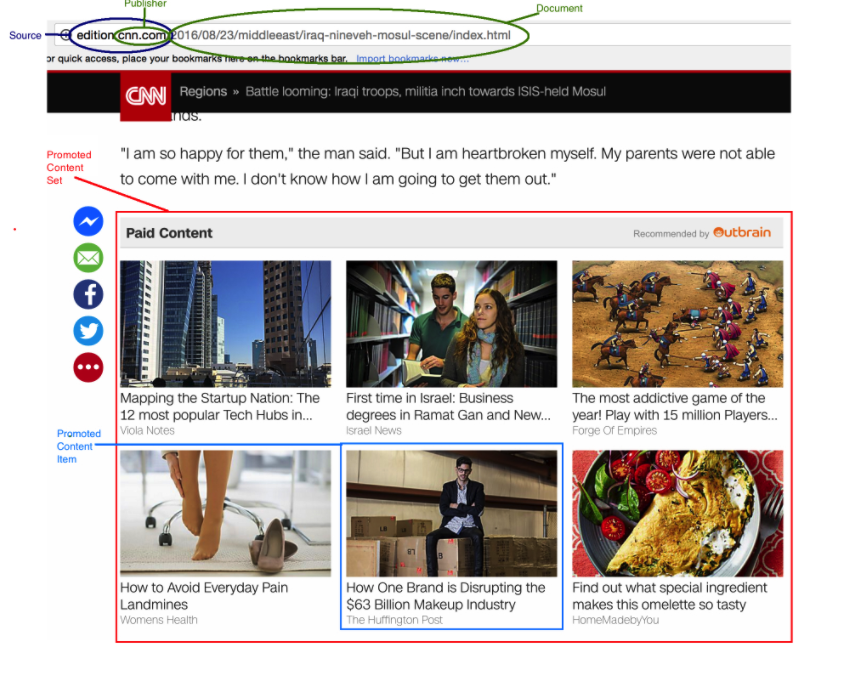
# **Data Set Description**

The dataset for this challenge contains a sample of users’ page views and clicks, as observed on multiple publisher sites in the United States between 14-June-2016 and 28-June-2016

Source - Kaggle

1. **page\_views.csv** is a log of users visiting documents. To save disk space, the timestamps in the entire dataset are relative to the first time in the dataset. If you wish to recover the actual epoch time of the visit, add 1465876799998 to the timestamp.
   1. uuid
   2. document\_id
   3. timestamp (ms since 1970-01-01 - 1465876799998)
   4. platform (desktop = 1, mobile = 2, tablet =3)
   5. geo\_location (country>state>DMA)
   6. traffic\_source (internal = 1, search = 2, social = 3)
2. **clicks\_train.csv** is the training set, showing which of a set of ads was clicked.
   1. display\_id
   2. ad\_id
   3. clicked (1 if clicked, 0 otherwise)
3. **clicks\_test.csv** is the same as clicks\_train.csv, except it does not have the clicked ad. This is the file you should use to predict. Each display\_id has only one clicked ad. Note that test set contains display\_ids from the entire dataset timeframe. Additionally, the public/private sampling for the competition is uniformly random, not based on time. These sampling choices were intentional, despite the possibility that participants can look ahead in time.
4. **events.csv** provides information on the display\_id context. It covers both the train and test set.
   1. display\_id
   2. uuid
   3. document\_id
   4. timestamp
   5. platform
   6. geo\_location
5. **promoted\_content.csv** provides details on the ads.
   1. ad\_id
   2. document\_id
   3. campaign\_id
   4. advertiser\_id
6. **documents\_meta.csv** provides details on the documents.
   1. document\_id
   2. source\_id (the part of the site on which the document is displayed, e.g. edition.cnn.com)
   3. publisher\_id
   4. publish\_time

**Outbrain as a publisher on CNN Website**



# **System Architecture:**

Model Deployment and working application

Model Design and predicting clicks

Feature Engineering

Meta Data Modelling

# **Models to be Used:**

Since we look at it as classification of clicks and predicting the probability of decreasing clicks we plan to use

* 1. Logistic Regression
  2. SVC

# **Languages and Deployment Tools:**

1. AWS/Azure
2. Docker
3. Python/R

# **Project Plan**

Documentation will take place at every stage

|  |  |
| --- | --- |
| **Dates** | **Project Task** |
| 17-04-2017 | Project Proposal |
| 18-04-2017 – 19-04-2017 | Metadata Modelling |
| 20-04-2017 – 21- 04 -2017 | Exploratory Analysis |
| 22-04-2017 – 25 – 04 - 2017 | Design of Model and Predicting clicks on Probabilities |
| 26-04-2017 – 27-04-2017 | Deployment /Testing/Revisiting |
| 28-04-2017 | Final Documentation and Git Uploads |
|  |  |

# **Contribution**

Equal contribution from all team members