A Datasheet for Dataset

1. Motivation
   (a) For what purpose was the dataset created? For research purposes.
   (b) Who created the dataset(e.g.,which team, research group) and on behalf of which entity (e.g., company, institution, organization)? Data is collected and provided by the publisher.
   (c) Who funded the creation of the dataset? No funding was involved; the publisher shared it with us for research purposes.
2. Composition
   (a) What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Instances of ads rendered on the users' browsers.
   (b) How many instances are there in total (of each type, if appropriate)? 2000 instances in the data sample.
   (c) Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? It's a data sample.
   (d) What data does each instance consist of? The dataset includes 30 anonymized features, along with a user ID and a binary label/target. Among these 30 features, there are 4 binary, 17 numerical, and 9 categorical features.
   (e) Is there a label or target associated with each instance? Yes, the target is whether the ad is viewed.
   (f) Is any information missing from individual instances? No.
   (g) Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? No.
   (h) Are there recommended data splits (e.g., training, development/validation, testing)? Yes, we did 80:10:10 for training, validation, and test sets respectively.
   (i) Are there any errors, sources of noise, or redundancies in the dataset? No.
   (j) Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? Yes, self-contained.
   (k) Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals' non-public communications)? No.
   (l) Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.
   (m) If the dataset does not relate to people, you may skip the remaining questions
      i. Does the dataset identify any sub populations (e.g., by age, gender)? No.
      ii. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? No.
      iii. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? No, the data sample is obfuscated to prevent any potential identifications.
3. Collection Process
   (a) How was the data associated with each instance acquired? The data was directly observable.
   (b) What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? Software programs run as JavaScript and some data were obtained from software APIs of the browser.
   (c) If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? deterministically providing 10 users data that each have 200 data points. These 200 Instances are selected while preserving the distribution of labels per user.
   (d) Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated(e.g., how much were crowdworkers paid)? An author was cooperating with publisher to obtain the data.
   (e) Over what timeframe was the data collected? Data sample provided is from the 30-days period and are collected in the month of November 2025.
   (f) Were any ethical review processes conducted (e.g., by an institutional review board)? Not applicable.
   (g) If the dataset does not relate to people, you may skip the remaining questions in this section.
      i. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? Not applicable
      ii. Were the individuals in question notified about the data collection? Not Applicable
      iii. Did the individuals in question consent to the collection and use of their data? Not applicable
      iv. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? Not applicable
      v. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? Not applicable
4. Preprocessing/cleaning/labeling
   (a) Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? Yes. We preprocessed the dataset by removing records with missing or null values, normalizing numerical

features, and encoding categorical ones. No extra labeling was needed since target values were already available for each instance.

(b) Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? Not applicable as we provide data sample.

(c) Is the software that was used to preprocess/clean/label the data available? Not applicable as we provide data sample.

5. Uses

(a) Has the dataset been used for any tasks already? No.

(b) Is there a repository that links to any or all papers or systems that use the dataset? No.

(c) What(other) tasks could the dataset be used for? Not Applicable as we are providing data sample.

(d) Is there anything about the composition of the dataset or the way it was collected and preprocessed/-cleaned/labeled that might impact future uses? Not Applicable as we are providing data sample.

(e) Are there tasks for which the dataset should not be used? Not Applicable as we are providing data sample.

6. Distribution

(a) Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? No.

(b) How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? Although not Applicable, data sample is shared in the GitHub repository mentioned in the paper.

(c) When will the dataset be distributed? Although not Applicable, the link to the repository containing the data sample is already included in the paper.

(d) Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? Not Applicable as we are providing data sample.

(e) Have any third parties imposed IP-based or other restrictions on the data associated with the instances? Not Applicable as we are providing data sample.

(f) Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? Not Applicable as we are providing data sample.

7. Maintenance

(a) Who will be supporting/hosting/maintaining the dataset? Not Applicable as we are providing data sample.

(b) How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Via authors email included in the paper.

(c) Is there an erratum? No.

(d) Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? Not Applicable as we are providing data sample.

(e) If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? Not Applicable

(f) Will older versions of the dataset continue to be supported/hosted/maintained? Not Applicable as we are providing data sample.

(g) If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Not Applicable as we are providing data sample.