Criteo Display Advertising Dataset Overview

This document provides an overview of the Criteo Display Advertising Challenge dataset. The dataset is widely used for building Click-Through Rate (CTR) prediction models. It consists of anonymized numerical and categorical features representing user behavior, ad characteristics, and contextual information. The target variable indicates whether an advertisement was clicked.

# Dataset Structure

The dataset consists of 40 columns:  
- 1 Label column (click / no-click)  
- 13 Numerical (integer/continuous) features (I1–I13)  
- 26 Categorical (hashed string) features (C1–C26)

# Column Description

|  |  |  |
| --- | --- | --- |
| Column Name | Type | Description |
| Label | Binary (0/1) | Target variable indicating whether the ad was clicked (1 = clicked, 0 = not clicked). |
| I1 – I13 | Numerical | 13 anonymized numerical features (counts, frequencies, or continuous values such as number of impressions, user statistics, or ad properties). Exact meanings are undisclosed for privacy. |
| C1 – C26 | Categorical (hashed) | 26 anonymized categorical features represented as hashed strings. These may include user demographics, device type, browser, ad category, advertiser ID, campaign ID, etc. |

# Link of the Dataset (Sample)

<https://drive.google.com/file/d/1ItAtlGuwvuV4LBNV5NbTP_Ep6VDalsHd/view?usp=sharing>

# Shape of the dataset:

Total rows: 45840617

Total columns: 40