# Predicting Website Ad Clicks

## Milestone: Final Project Proposal

Group 16

Amitesh Tripathi
Sayali Lad

857 3761991(Tel:Student 1)
857 277 4326(Tel:Student 2)

tripathi.am@northeastern.edu
lad.sa@northeastern.edu

**Percentage of Effort Contributed by Student 1: 50 %**

**Percentage of Effort Contributed by Student 2: 50 %**

**Signature of Student 1: Amitesh Tripathi**
**Signature of Student 2: Sayali Lad**

**Submission Date: 02/03/2023**

# Final Project Proposal: Predicting Website Ad Clicks

# IE 7275: Data Mining in Engineering

**Problem Setting:**
An e-commerce website is looking to improve sales through targeted advertisements on partner websites. The website has hired an Adtech company to build a system for displaying ads for products that customers have previously viewed or similar items. The goal is to predict the probability of a user clicking on an ad based on their viewing history and user data.

**Problem Definition:**
The task is to predict the likelihood that a user clicks on a product ad on a partner website. This will be done by analyzing the user's view log, ad impression, and user data to determine the probability of a click in the next 7 days.

**Data Sources:**
The data for this project comes from the e-commerce website and includes view log data from October 15, 2018, to December 11, 2018, product description data, and ad impression data from November 15, 2018, to December 18, 2018. The data includes train and test sets, with the train set containing information on ad impressions and whether or not the ad was clicked. The test set contains ad impression information without labels.

https://www.kaggle.com/datasets/arashnic/ctrtest?resource=download2019/#ProblemStatement
https://www.kaggle.com/datasets/jahnveenarang/cvdcvd-vd
https://github.com/splikhita/Ad-Click-Prediction/blob/master/Advertisements-Data.csv

**Data Description:**
The column names, obtained from multiple datasets, play a crucial role in the consolidation process as they will be utilized to merge the datasets into a single, unified dataset for further data exploration and analysis. The chosen columns will be carefully selected based on their relevance to the goals of the analysis, ensuring the quality and consistency of the final dataset.

Following is the column name and the description of it.

**Column Nam**e: Description
**Age**: The age of the individual
**Daily Time Spent on Site**: The amount of time spent on the website
**City**: The city where the individual resides
**Country**: The country where the individual resides
**Gender**: Gender of the individual

**EstimatedSalary**: Estimated salary of the individual
**Daily Internet Usage**: Daily internet usage of the individual
**Ad topic line**: The topic of the marketing ad
**Clicked on Ad**: Indicates whether the individual clicked on the ad (1 for yes, 0 for no)
**Impression_id**: Unique identifier for ad impression
**Impression_time**: The time when the ad was shown to the individual
**User_id**: Unique identifier for the individual
**Os_version**: Operating system version of the individual's device
**Is_4G**: Indicates whether the individual's device is using 4G (1 for yes, 0 for no)
**Is_click**: Indicates whether the individual clicked on the ad (1 for yes, 0 for no)
**Server_time**: The time when the ad was served by the server
**Device_type**: Type of device used by the individual
**Session_id**: Unique identifier for the individual's session
**Item_id**: Unique identifier for the item advertised