# milestone_2_00951537

June 13, 2024

## 1 Introduction

[ ]:

```
# Preamble
## Imports
import pandas as pd
```

The global digital advertising market is worth approximately $602 billion today. Due to the increasing rate of of online participation since the COVID-19 pandemic, this number has been rapidly increasing and is expected to reach $871 billion by the end of 2027 (eMarketer, 2024). Many of the of the major Ad platforms such as Google, Facebook and Amazon operate on a cost-per-user-engagement pricing model, which usually means that advertisers get charged for every time a user clicks on an advertisment. This means that these platforms are incentivized to make sure that the content shown to each user is as relevent as possible in order to maximize the number of clicks in the long term. Attaining accurate Click-Through Rate (CTR) prediction is a necessary first step for Ad personalization, which is why study of CTR prediction methods have been an extremely active part of Machine Learning research over the past through years.

Initially, shallow prediction methods such as XGBoost (Cite), Factorization Machines (Cite) and Field-Aware Factorization Machines (Cite) have been used for CTR prediction. However, these methods have often been shown to be unable to capture the higher order feature interactions in the sparse multy value categorical Ad Marketplace datasets (Cite). Since then, Deep Learning methods have been shown to show superior predictive ability on these datasets. The focus of my reasearch project is therefore to explore the merits of different Deep Learning architechtures for click-through rate prediction

In the following report, I explore the relevant datasets and simulations that I will be using throughout my research project. In the first section, I perform an exploratory data analysis on three widely adopted benchmark CTR prediction datasets; the KDD12 (Aden, 2012), Avazu (Wang and Cukierski, 2014) and Criteo (Tien et al, 2014) datasets. In the second section, I then explore possible ways of simulating the ad marketplace environment in order to test the reinforcement learning framework.

## 2 Datasets

### 2.1 KDD12

The **KDD12** dataset was first released for the KDD Cup 2012 competition (Cite), with the original task being to predict the number of clicks for a given number of impressions. Each line represents a training instance derived from the session logs for the advertizing marketplace. In the context of this dataset, a “session” refers to an interaction between a user and the search engine, containing the following components; the user, a list of adverts returned by the search engine and shown (impressed) to the user and zero or more adverts clicked on by the user. Each line in the training set includes:

- **Clicks**: The number of times the user has clicked on the given Ad among the relevant Impressions.
- **Impressions**: The number of search sessions in which the Ad was impressed by the user after issuing the specific Query.
- **Display URL**: The URL link displayed along with the advert.
- **Ad ID**: An identifier for each advertisment.
- **Advertiser ID**: An identifier that specified the company that issued the advertisment.
- **Depth**: The number of Ads that the user viewed (impressed) in a session.
- **Position**: The order in which the specific advert was displayed to the user within the session.
- **Query ID**
- **Keyword ID**
- **Title ID**
- **Description ID**
- **User ID**

Describe each of the datasets

```
[ ]: # Show firt 5 rows of the training dataset
     data = pd.read_csv('.\data\kdd12\kdd12_training.csv')
     data.head()
```

```
[ ]:    Click  Impression            DisplayURL      AdID  AdvertiserID  Depth  \
     0      0           1   4298118681424644510   7686695           385      3
     1      0           1   4860571499428580850  21560664         37484      2
     2      0           1   9704320783495875564  21748480         36759      3
     3      0           1  13677630321509009335   3517124         23778      3
     4      0           1   3284760244799604489  20758093         34535      1

        Position  QueryID  KeywordID  TitleID  DescriptionID  UserID
     0         3     1601       5521     7709            576  490234
     1         2  2255103        317    48989          44771  490234
     2         3  4532751      60721   685038          29681  490234
     3         1     1601       2155     1207           1422  490234
     4         1  4532751      77819   266618         222223  490234
```

### 2.2 Avazu

### 2.3 Criteo

## 3 Simulatinons

## 4 Summary of findings

## 5 Suggested Future Research

## 6 References

- eMarketer. (2023). Digital advertising spending worldwide from 2021 to 2027 (in billion U.S. dollars) . Statista. Statista Inc.. Accessed: June 09, 2024. https://www-statista-com.iclibezp1.cc.ic.ac.uk/statistics/237974/online-advertising-spending-worldwide/
- Aden, Yi Wang. (2012). KDD Cup 2012, Track 2. Kaggle. https://kaggle.com/competitions/kddcup2012-track2
- Steve Wang, Will Cukierski. (2014). Click-Through Rate Prediction. Kaggle. https://kaggle.com/competitions/avazu-ctr-prediction
- Jean-Baptiste Tien, joycenv, Olivier Chapelle. (2014). Display Advertising Challenge. Kaggle. https://kaggle.com/competitions/criteo-display-ad-challenge