Capstone Project

Wilson Chen

Machine Learning Engineer Nanodegree

September 7, 2020

Domain Background

Digital marketing has became the main channel companies advertising their products and services. People now heavily rely on internet to work and to live then ever. From grocery delivery apps to video platforms, there are thousands of places to display ads. Also, digital marketing ecosystem is growing more and more complex while new platform and technologies rise. Advertisers are facing challenge on how to efficiently advertise their ads. The ideal is inspired by this article on The New York Times.

Problem Statement

The challenges digital marketing facing are:

1. People hate to see ads they are not interested in or irrelevant. When seeing such ads, they will skip it or even have negative response to the brand or products. This is a cost to the advertiser.

2. While new digital channels like podcast emerge and ads fee rises, advertiser need a measure to evaluate their digital marketing results to decide how to spend the budget or design the ads.

Solution Statement

The goal is to create a Click Through Rate(CTR) prediction model and deploy it to a website for users to use. This model can help:

1. Display ads to right audience who are interested in and will click through the ads.

2. Provide a measure to predict an ad’s performance to advertisers.

Data

2020 DIGIX Advertisement CTR Prediction provided by Louis Chen7 on Kaggle.com

This data has advertising behaviors collected from seven consecutive days recording whether there is a click or not. Note I only use train_data.csv because there is no label in test_data.csv.

Benchmark Model

Since this is a classification problem. A simple logistic regression model is a good benchmark model.

Evaluation Metrics

The definition of CTR is the ratio of users who click on an ad to the number of total users who see the ad.

$$\text{CTR} = \frac{\#\ of\ \ users\ \ click\ \ the\ \ ad}{\#\ of\ \ users\ \ see\ \ the\ \ ad}$$

Because we can control what user can see, this can be interpreted as follow:

$$\text{CTR} = \frac{\#\ of\ \ users\ \ click\ \ the\ \ ad}{\#\ of\ \ users\ \ the\ \ we\ \ show\ \ the\ \ ad}$$

This is actually the precision from predicting perspective.

$$\text{Precision} = \frac{\#\ of\ users\ click\ the\ ad}{\#\ of\ users\ the\ model\ shows\ the\ ad}$$

Therefore, I choose the precision as the metric to evaluate the model performance.

Project Design

To achieve the goal, the workflow will include following steps:

1. Read at least two CTR prediction papers to learn more about the topic and avoid known issues that are worthless to invest my time.

2. Explore the data and visualize the findings.

3. Perform feature engineering to create new or select essential features to work on.

4. Create the benchmark logistic regression classification models for later evaluation.

5. Develop classification models and choose the model with highest precision as final model.

6. Deploy the model to an endpoint.

7. Create a webpage allowing users to view analysis visualizations and interact with final model.