CTR Prediction

Wilson Chen

Machine Learning Engineer Nanodegree Capstone Project

September 7, 2020

Domain Background

Digital marketing has become the primary channel companies advertising their products and services. People now heavily rely on the internet to work and to live than ever. From grocery delivery apps to video platforms, there are thousands of places to display ads. Also, the digital marketing ecosystem is growing more and more complex since new technologies bring new channels such as live streaming. Advertisers are facing challenges on how to advertise their ads efficiently. This article on The New York Times was the project inspiration.
Since there are more and more un-skippable ads, I, as a heavy internet user, want to see more relevant content as well. I am very interested in improving my online experience.

Problem Statement

The challenges digital marketing facing are:

1. People hate to see ads they are not interested in or irrelevant. When getting such ads, they will skip it or even have negative responses to the brand or products.
2. While new digital channels like podcasts emerge and ads fee rises, advertisers 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 the 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

I will use 2020 DIGIX Advertisement CTR Prediction [1] 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. It contains a label that indicates whether an ad is clicked along with features from users, apps, and devices.

I will only use train_data.csv because there is no label in test_data.csv. The whole data file is about 6 GB. Since this file is too large to push to Github, I will include instructions on how to download the data in README. Also, because the size of the data is too huge for this project, I will conduct a stratified sampling to get a subset of data to work on it.

Benchmark Model

According to the paper Ad Click Prediction: a View from the Trenches [2], the logistic regression performs well in ads click prediction. In the research paper, they use FTRL-Proximal with logistic regression instead of stochastic gradient descent because of memory usage problems. Since I am only comparing the model performance, I will use general stochastic gradient descent with logistic regression.
The definition of benchmark logistic regression is as follow:

We denote X to be the feature vector, θ to be the coefficient vector, and $z = \theta^T \cdot X$

Then the logistic regression function will be $h_\theta(X) = \frac{1}{1+e^{-z}}$

And then we assign y to be the actual label, the cost function to minimize is
$J_\theta = -y \cdot \log\left(h_\theta(X)\right) - (1-y) \cdot \log\left(1 - h_\theta(X)\right)$

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, the denominator can convert to:

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

Then we find out that CTR is 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

The workflow will include the following steps:

1. Explore the data and visualize the findings.
2. Perform feature engineering to create new or select essential features for models.
3. Create the benchmark logistic regression classification models for later evaluation.
4. Develop classification models and choose the model with the highest precision as the final model.
5. Deploy the model to an endpoint.
6. Create a webpage allowing users to view analysis visualizations and interact with the final model.

Reference

[1] Louis Chen7 2020 DIGIX Advertisement CTR Prediction Kaggle
https://www.kaggle.com/louischen7/2020-digix-advertisement-ctr-prediction
[2] H. Brendan McMahan, Gary Holt, D. Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, and Jeremy Kubica Ad Click Prediction: a View from the Trenches Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (2013)
https://research.google/pubs/pub41159/