캐글 모델

1등 솔루션

TalkingData AdTracking Fraud Detection Challenge

Can you detect fraudulent click traffic for mobile app ads?

https://www.kaggle.com/competitions/talkingdata-adtracking-fraud-detection/discussion/56475

I. 다운 샘플링

- is_attributed 의 불균형 0/1 → 99.8 % / 0.2%
- is_attributed 이 0인 데이터 샘플을 축소시켜도 성과에서 큰 문제가 없었다는 점 관찰.

II. 피쳐 엔지니어링

- 다운 샘플링 대신 이 전략을 통해 모든 데이터를 사용
- 사용한 변수
  - 기본 제공된 변수: ip, os, app, channel, device
    - 31 (=(2^5) - 1)개의 조합을 만들고, 각각의 조합에 대해 “click 시리즈 기반”의 특성 세트를 생성 → full_click_train.csv

    ex) 다음 1시간 or 6시간 내의 click 수, 앞뒤 click 시간 간격, **과거 클릭들의 평균 attributed 비율**

  - 시간 관련 변수: day, hour

→ 따로 피쳐 selection는 안하고, 위의 변수를 모두 넣어 학습. 이 때 LGB 모델의 스코어는 0.9808

III. LDA/NMF/LSA를 이용한 categorical feature embedding

- categorical feature embedding → categorical feature를 머신 러닝에서 모델이 학습할 수 있도록 변수를 적절하게 가공하는 작업?

- LDA (latent Dirichlet allocation)
  - 차원 축소 기법

```
apps_of_ip = {}
for sample in data_samples:
  apps_of_ip.setdefault(sample['ip'], []).append(str(sample['app']))
ips = list(apps_of_ip.keys())
apps_as_sentence = [' '.join(apps_of_ip[ip]) for ip in ips]
apps_as_matrix = CountTokenizer().fit_transform(apps_as_sentence)
topics_of_ips = LDA(n_components=5).fit_transform(apps_as_matrix)
```

  5개의 기초 변수(ip, os, app, channel, device)으로 가능한 모든 조합(5 * (5 - 1) = 20가지)에 대해 LDA를 사용하여 주제를 추출하고, 각 주제의 크기를 5로 설정 → 총 100개의 피쳐 생성

- NMF, PCA도 비슷한 방식으로 피쳐 생성 → 총 300개의 피쳐 생성
- 이후 app 변수만 남기고 **나머지 4개 기초 변수 제거** → LGBM 모델 기준, 성과 향상 (0.9821 → 09828)

  → **one-hot encoding**

IV. NN 모델

- 3층의 모델
- LGB 모델보다 조금 낮은 성과
- “성과를 올리기 위해서는 is_attributed 가 0인 데이터가 더 필요해보였다”

V. 검증

We used day 7 & 8 for training and day 9 for validation, and chose the best number of iterations of LGB. Then, we trained a model on day 7 & 8 & 9 with the obtained number of iterations for creating submission.

→ 7일, 8일 데이터로 학습, 9일 데이터로 검증하여 모델 선

VI. 최종 제출

- rank-based weighted averaging - 성과에 따라 랭킹을 부여하고 가중치를 부여
    - 7개의 LGBM 모델 + 1개의 NN 모델 (0.98343)

3등 솔루션(RNN)

TalkingData AdTracking Fraud Detection Challenge
Can you detect fraudulent click traffic for mobile app ads?
https://www.kaggle.com/competitions/talkingdata-adtracking-fraud-detection/discussion/56262

- the click delta is important,so I fed deltas of last 5 and next 5 click_times to the network and designed a model with RNN cell to find the patterns of the click series

6등 솔루션

TalkingData AdTracking Fraud Detection Challenge
Can you detect fraudulent click traffic for mobile app ads?
https://www.kaggle.com/competitions/talkingdata-adtracking-fraud-detection/discussion/56283

it is clear now that the test data was sorted by click time then target value

→ 테스트 데이터가 클릭 시간, 그 다음 타겟 값을 기준으로 정렬됨

관찰

- ip, app, device, os, channel 각각의 nunique 개수는 많지만, 조합으로 고려했을 때는 겹치는 것이 많다.

- test데이터의 행은, `click_time` 을 기준으로 정렬된 상태로 제공

코랩 pro

- 대용량 데이터를 다루는 방법
  - RAM
  - 데이터 타입 정의(uint)
  - DASK 사용하기 → 아직 안 이용함

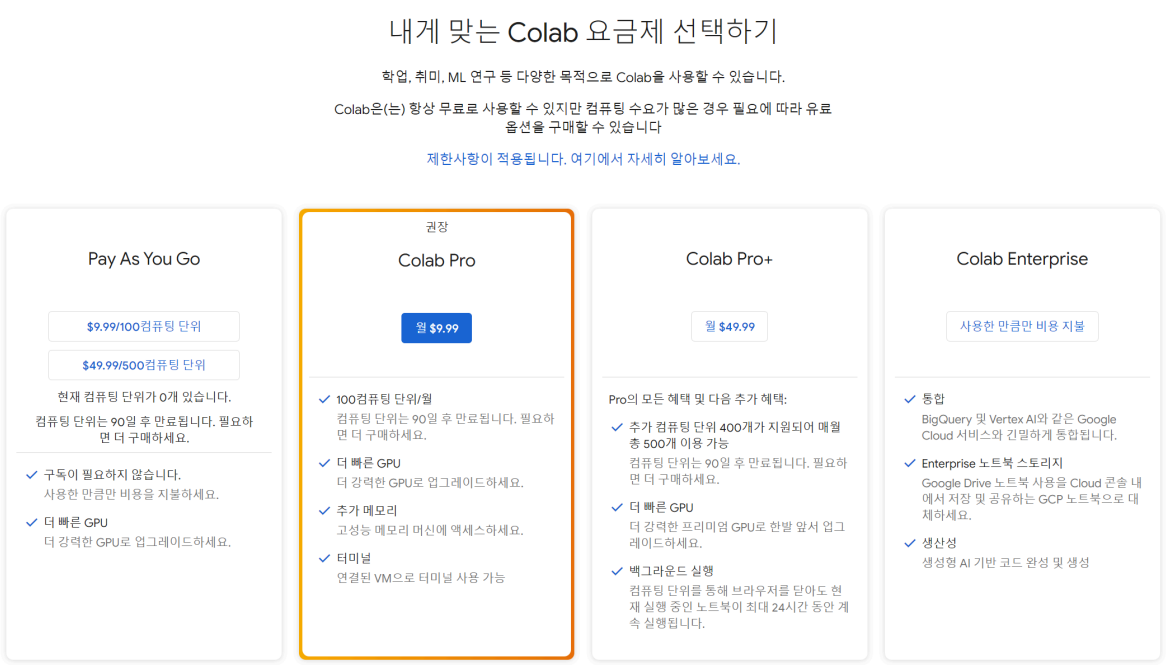
```
import dask.dataframe as dd

dask_df = dd.read_csv('large_file.csv')

# Dask 데이터프레임에 대해 필요한 작업 수행
dask_df = dask_df.groupby('ip').rolling('1h', on='click_time').count().co
```

Colab Paid Services Pricing

https://colab.research.google.com/signup



- ip을 제외한 조
- K-Fold : 시간을 고려하도록 수정 (Non-stationary)
  - day 7 & day 8 & day9
    - 1 2 3 → 4
    - 2 3 4 → 5

  - 목표 : 답지를 만드는