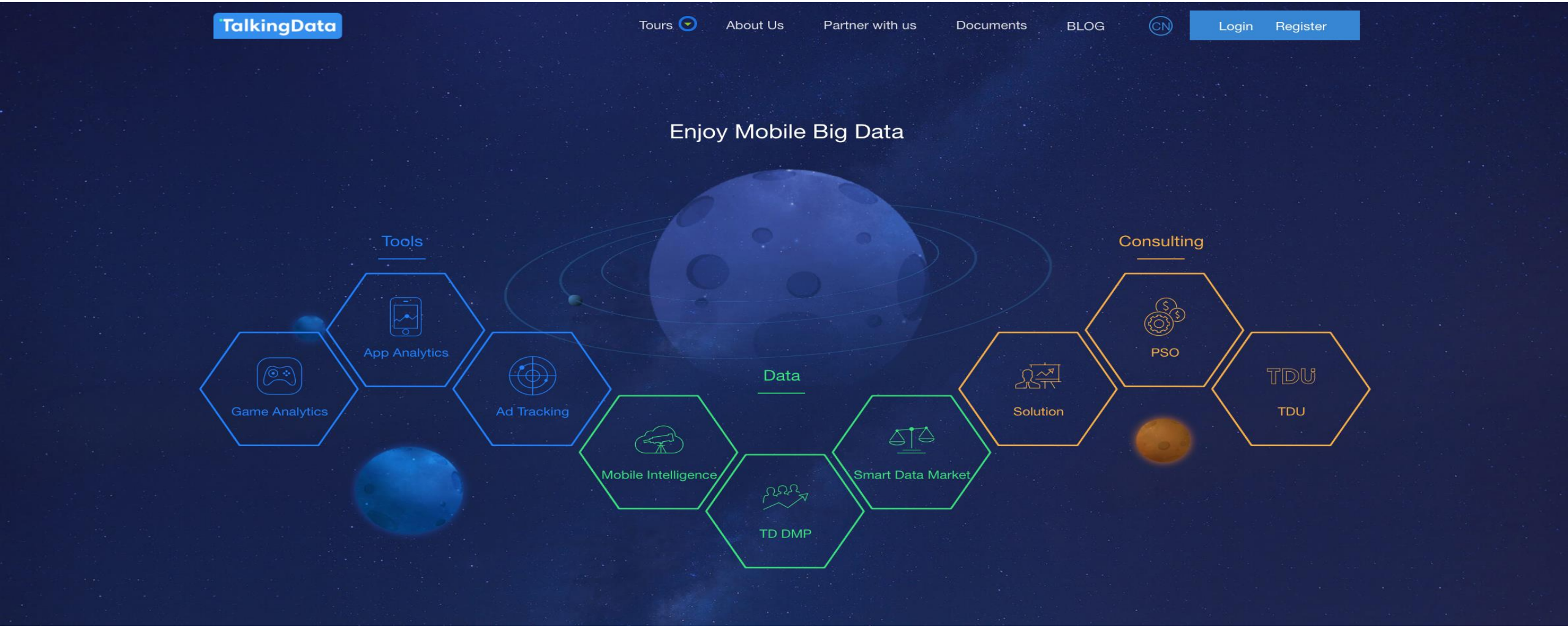
TalkingData AdTracking Fraud Detection Challenge

Can you detect fraudulent click traffic for mobile app ads?

Index

Introduction



Introduction

- 광고를 맡긴 회사가 원하는 것

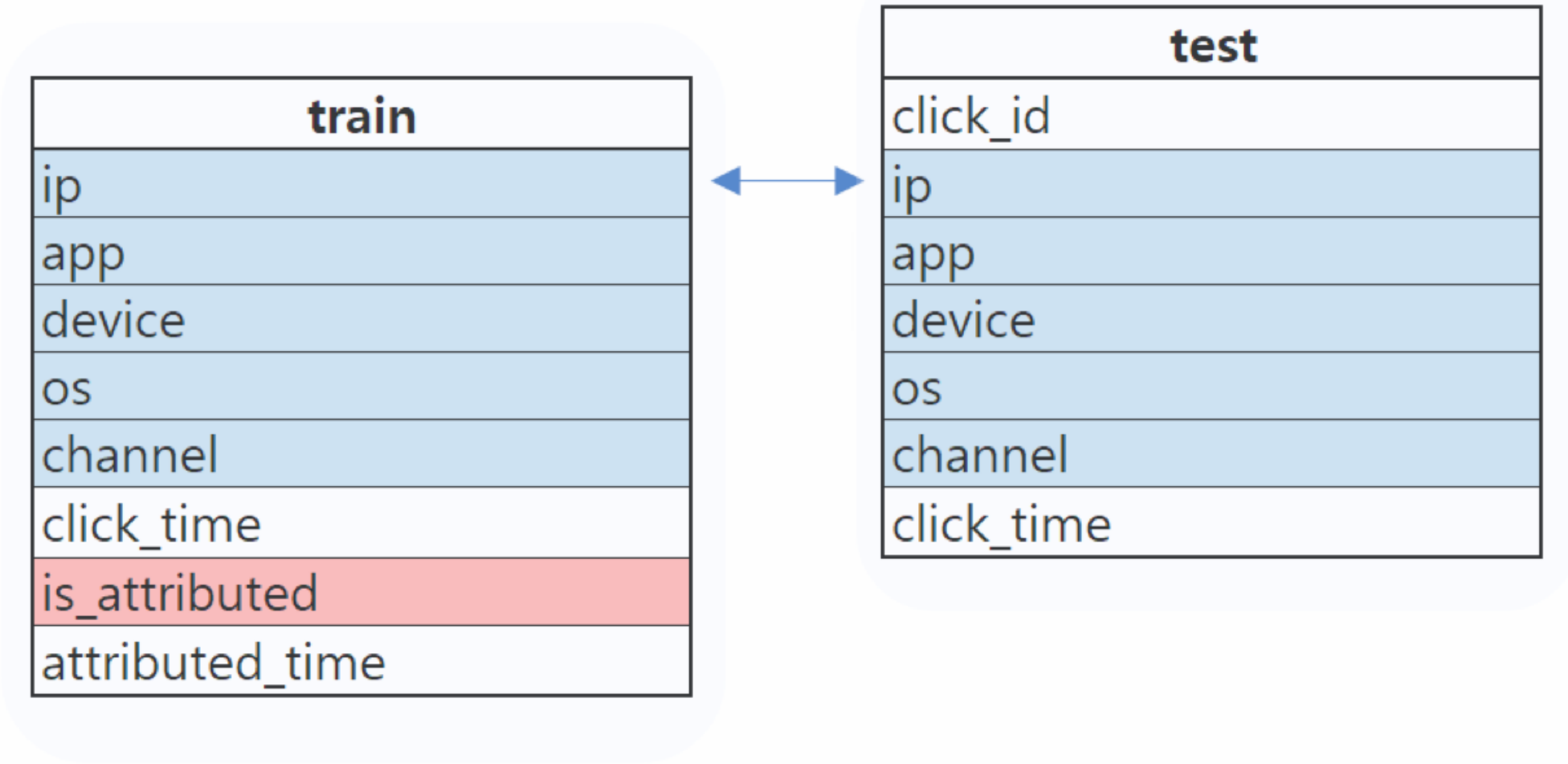
광고를 보는 사람들이 광고를 클릭하고 그 앱을 다운로드 받는 행위까지 연결되는 것.

- 그러나, 광고를 클릭만 해도 광고를 맡긴 회사를 돈을 내야 한다.

- 광고를 클릭하기만 하고 다운 받지 않는 사람들을 사기라고 간주.

- 즉, 사기치는 사람들을 골라내자!

Data Description



Data Description

train
ip
app
device
os
channel
click_time
is_attributed
attributed_time

train
ip
app
device
os
channel
day
hour
is_attributed

Data Description

train
ip
app
device
os
channel
day
hour
is_attributed

group_by
['ip'], 'channel'
['ip'], 'os'
['ip'], 'device'
['ip', 'day'], 'hour'
['ip', 'day', 'channel']
['ip', 'app', 'channel'], 'day'
['ip', 'app', 'channel'], 'hour'
.
.
.
is_attributed

Data Description

train
ip
app
device
os
channel
day
hour
is_attributed

group_by
['ip'], 'channel'
['ip'], 'os'
['ip'], 'device'
['ip', 'day'], 'hour'
['ip', 'day', 'channel']
['ip', 'app', 'channel'], 'day'
['ip', 'app', 'channel'], 'hour'
.
.
.
is_attributed

Data Description

group_by
['ip'], 'channel'
['ip'], 'os'
['ip'], 'device'
['ip', 'day'], 'hour'
['ip', 'day', 'channel']
['ip', 'app', 'channel'], 'day'
['ip', 'app', 'channel'], 'hour'
.
.
.
is_attributed

def
- count
- countuniq
- cumcount
- mean
- var
- prevclick
- nextclick
- test_hh

Feature Engineering PrevClick

ip	os	device	channel	click_time
1	2	1	43	2017-11-07 9:30
1	2	1	132	2017-11-07 13:40
1	2	1	43	2017-11-07 18:05
1	2	1	132	2017-11-07 4:58
1	2	1	43	2017-11-09 9:00
1	2	1	132	2017-11-09 1:22
1	2	1	43	2017-11-09 1:17
1	2	1	132	2017-11-07 10:01

ip_channel_prevClick
nan
nan
30900.0
58800.0
53700.0
69720.0
58620.0
31140.0

Feature Engineering NextClick

ip	os	device	channel	click_time	device_channel_nextClick
1	2	1	43	2017-11-07 9:30	30900.0
1	2	1	132	2017-11-07 13:40	58800.0
1	2	1	43	2017-11-07 18:05	53700.0
1	2	1	132	2017-11-07 4:58	69720.0
1	2	1	43	2017-11-09 9:00	58620.0
1	2	1	132	2017-11-09 1:22	31140.0
1	2	1	43	2017-11-09 1:17	nan
1	2	1	132	2017-11-07 10:01	nan

Feature Engineering Count

ip	os	device	channel	day	hr
1	2	1	43	7	13
1	2	1	132	7	14
1	2	1	43	7	15
1	2	1	132	8	12
1	2	1	43	8	12
1	2	1	132	9	14
1	2	1	43	9	15
1	2	1	132	9	16

ip_oscount
8
8
8
8
8
8
8
8

Feature Engineering Countuniq

ip	os	device	channel	day	hr	ip_by_channel_countuniq
1	2	1	43	7	13	2
1	2	1	132	7	14	2
1	2	1	43	7	15	2
1	2	1	132	8	12	2
1	2	1	43	8	12	2
1	2	1	132	9	14	2
1	2	1	43	9	15	2
1	2	1	132	9	16	2

Feature Engineering Cumcount

ip	os	device	channel	day	hr
1	2	1	43	7	13
1	2	1	132	7	14
1	2	1	43	7	15
1	2	1	132	8	12
1	2	1	43	8	12
1	2	1	132	9	14
1	2	1	43	9	15
1	2	1	132	9	16

ip_by_os_cumcount
0
1
2
3
4
5
6
7

Feature Engineering Mean

ip	os	device	channel	day	hr	ip_os_by_hr_mean
1	2	1	43	7	13	13.875
1	2	1	132	7	14	13.875
1	2	1	43	7	15	13.875
1	2	1	132	8	12	13.875
1	2	1	43	8	12	13.875
1	2	1	132	9	14	13.875
1	2	1	43	9	15	13.875
1	2	1	132	9	16	13.875

Feature Engineering Var

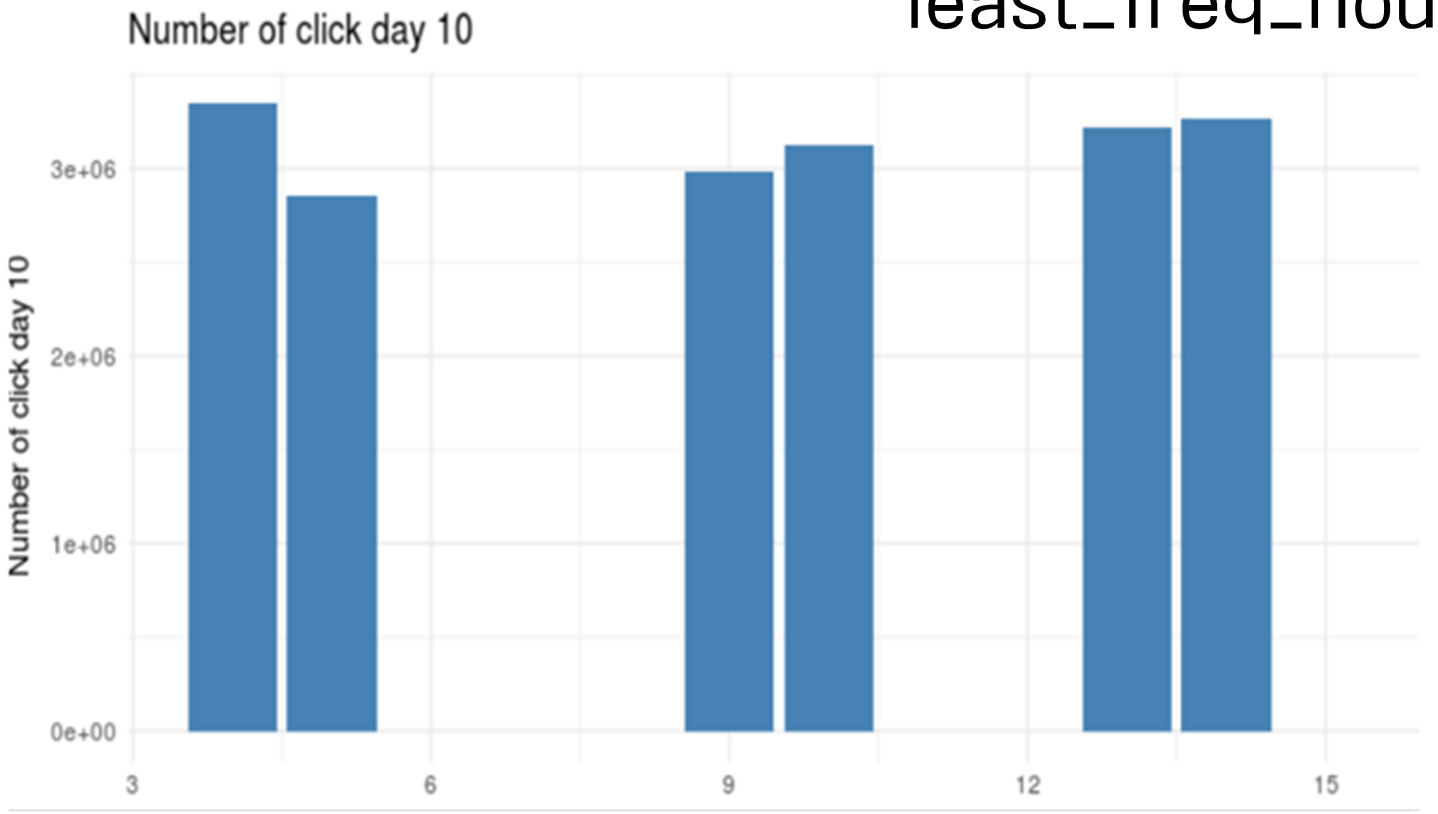
ip	os	device	channel	day	hr	ip_os_by_hr_var
1	2	1	43	7	13	2.125
1	2	1	132	7	14	2.125
1	2	1	43	7	15	2.125
1	2	1	132	8	12	2.125
1	2	1	43	8	12	2.125
1	2	1	132	9	14	2.125
1	2	1	43	9	15	2.125
1	2	1	132	9	16	2.125

Feature Engineering Test_hh

ip	os	device	channel	hour	day
1	2	1	43	9	7
1	2	1	132	13	7
1	2	1	43	18	7
1	2	1	132	6	7
1	2	1	43	9	9
1	2	1	132	1	9
1	2	1	43	1	9
1	2	1	132	10	7

Feature Engineering Test_hh

most_freq_hours = -1 * [4, 5, 9, 10, 13, 14]
least_freq_hours = +1 * [6, 11, 15]



Feature Engineering Test_hh

ip	os	device	channel	hour	day
1	2	1	43	9	7
1	2	1	132	13	7
1	2	1	43	18	7
1	2	1	132	6	7
1	2	1	43	9	9
1	2	1	132	1	9
1	2	1	43	1	9
1	2	1	132	10	7

in_test_hh
1
1
2
3
1
2
2
1

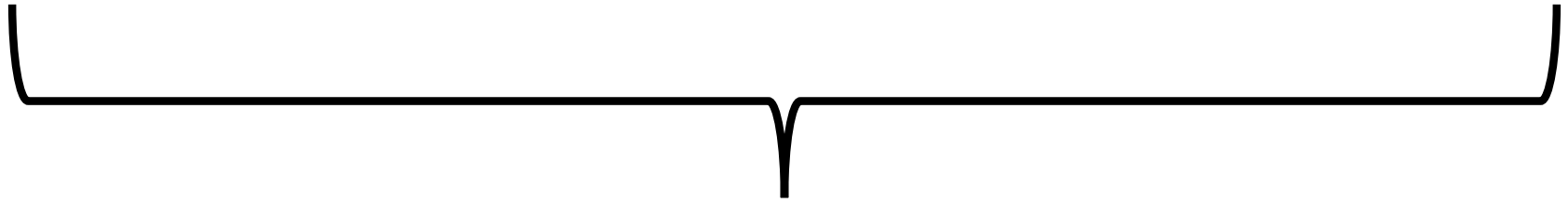
Result

Lightgbm

0.9800

Catboost

0.9785



기하· 조화평균 : 0.9800

Feedback (대회 전)

1. 중간고사 기간이어서 시간 분배가 힘들었다.
    - 시작은 3명이었는데 끝은 2명이었다.
2. 데이터가 너무 커서 가공 및 모델 돌리는데 시간이 오래 걸렸다.
    - 캐글 커널에서 잘 안돌아간다.(6시간 이상 안 돌아감/ 메모리도 부족)
    - 2500만개 까지 캐글커널에서 가능하다.
3. parameter tuning 시도는 해봤는데, 한계가 많다.
4. 변수생성 아이디어를 떠올리기 힘들다.
5. 재미가 없다!!!

Feedback (대회 후)

- 너무 크다보니깐 downsample로 빠른 모델가공하고 빠르게 결과값을 뽑아봤으면 좋았다.
- Auc를 기준으로 하는거여서 0과 1을 맞추는건데 우린 확률값으로 해서 진짜 0같은애들은 0으로 맞추고 1인 애들은 1로 맞춰주면 올랐다.
- Test supplement를 사용하면 실제로 결과가 올랐다.
- 쿼리이용해서 빠른속도로 모델 가공을 한 사람도 있었다.
- Xgboost gbdt말고 다트를 사용했어도 결과가 좋았다.
- 9일을 validation으로 사용해서 test해봤어도 좋았다.
- 상위권은 nn도 사용했다.
- 이 대회에서 사용된 방법은 wordbatch랑 nn 두가지가 있었다.

Thank you

Any questions or suggestions?