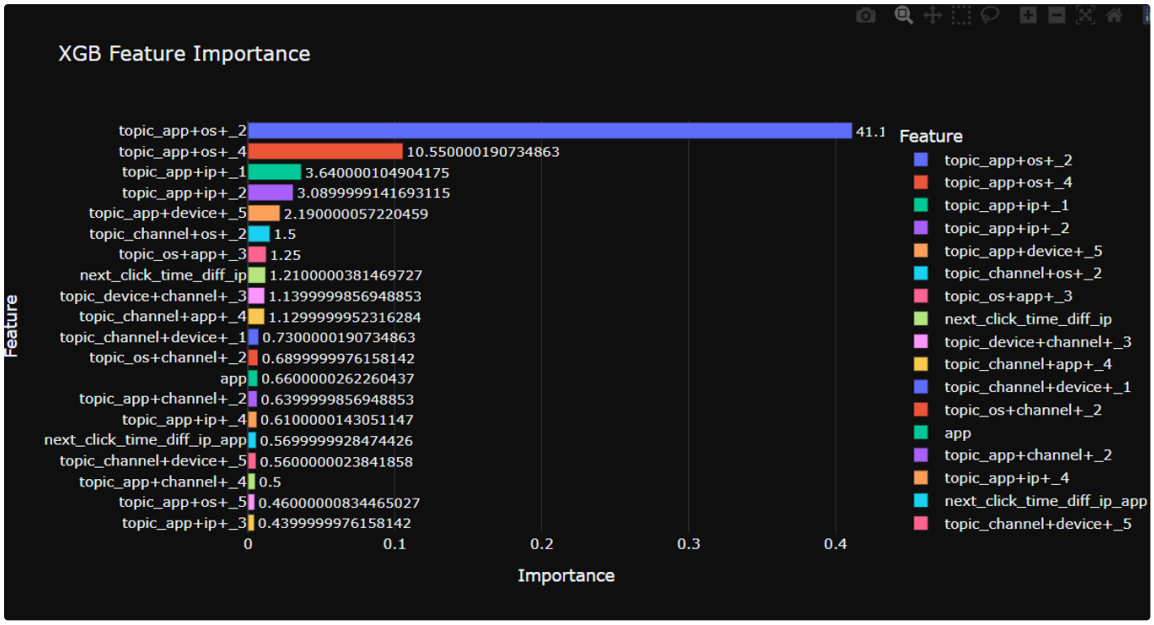
08.07~

- test 버리기
  - time_dh 불연속적이므로, test 데이터를 버리는 것으로
- Kmeans

- LDA 100개 중에 골라서
  - 피쳐 중요도 상위 몇 개만 남겨두는 방식



  - **app X 4개 (+20개)**
  - **4개 X app (+20개)**
  - channel X 4개
  - 4개 X channel

⇒ 40개만 채택

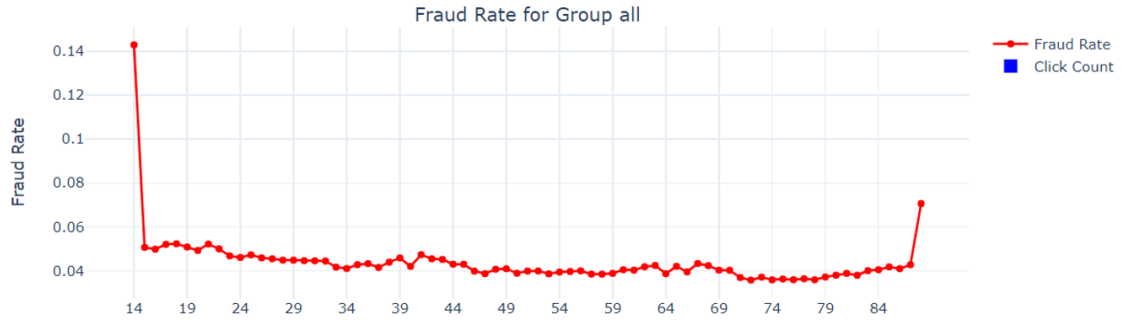
  - threshold

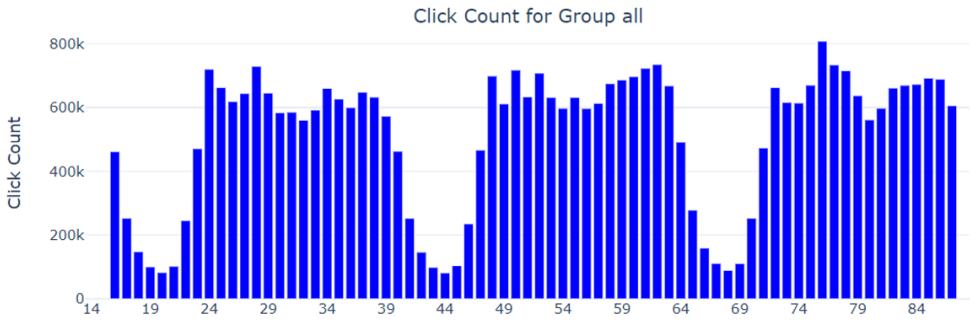
    = 0.000034

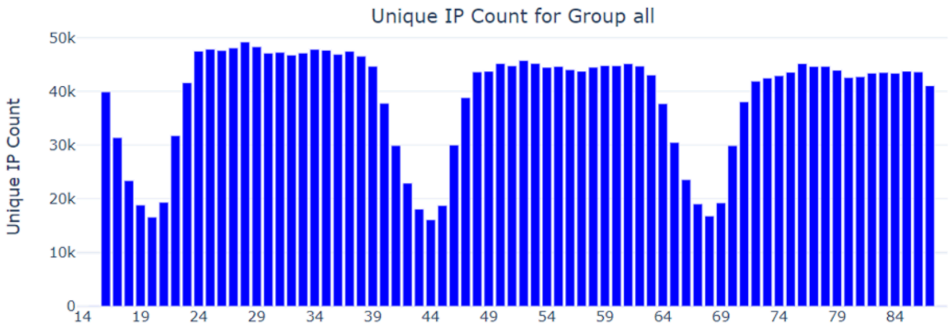
```
10261 4.971896501598992%
```

    - 전체 IP유저: `206,380`

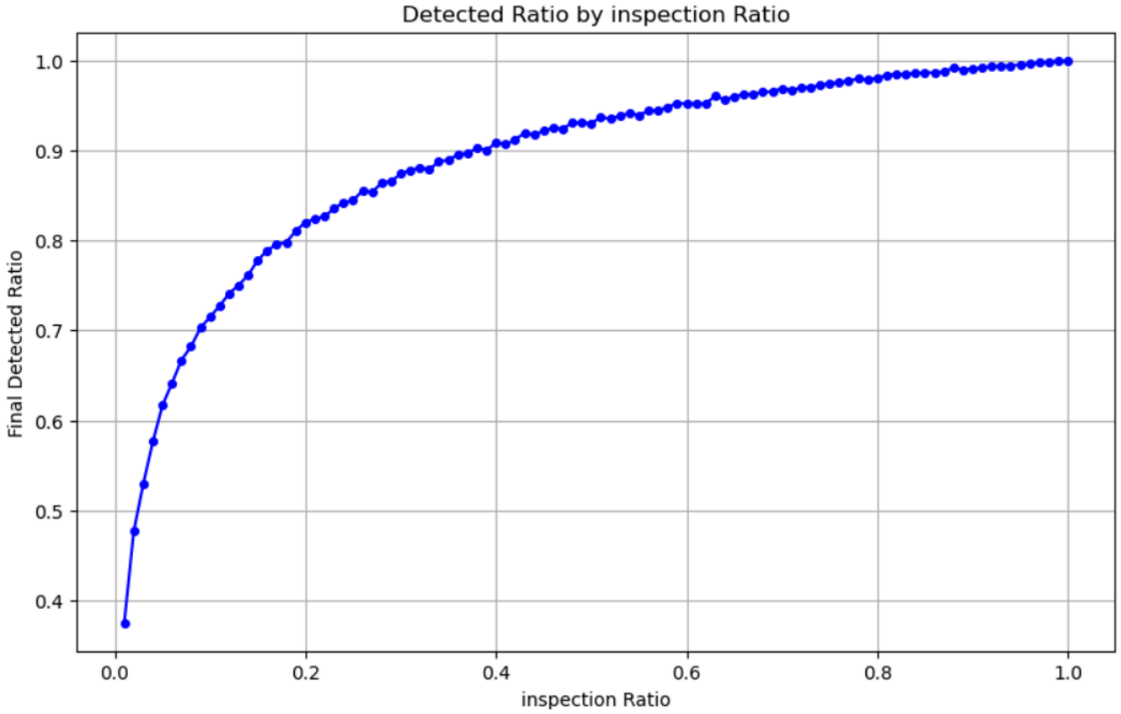
Group all


Fraud Rate for Group all
Fraud Rate
Click Count
Fraud Rate
0.14
0.12
0.1
0.08
0.06
0.04
14 19 24 29 34 39 44 49 54 59 64 69 74 79 84


Click Count for Group all
Click Count
800k
600k
400k
200k
0
14 19 24 29 34 39 44 49 54 59 64 69 74 79 84


Unique IP Count for Group all
Unique IP Count
50k
40k
30k
20k
10k
0
14 19 24 29 34 39 44 49 54 59 64 69 74 79 84

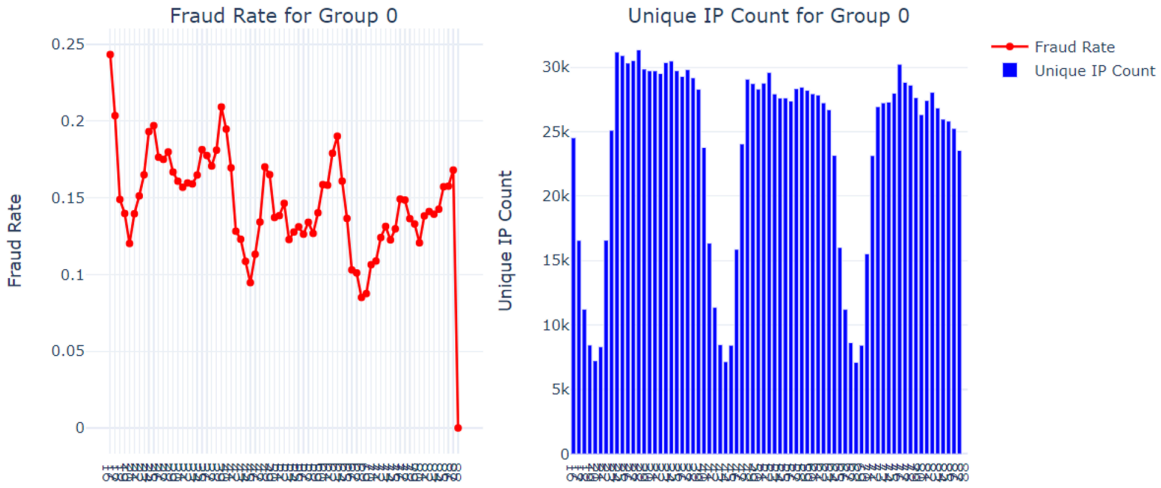
- 한 번 걸러진 IP는 계속 블락
  - y축 : 모든 시간이 흘렀을 때, 최종적으로 detect한 fraud 유저 비율
  - x축: inspection rate



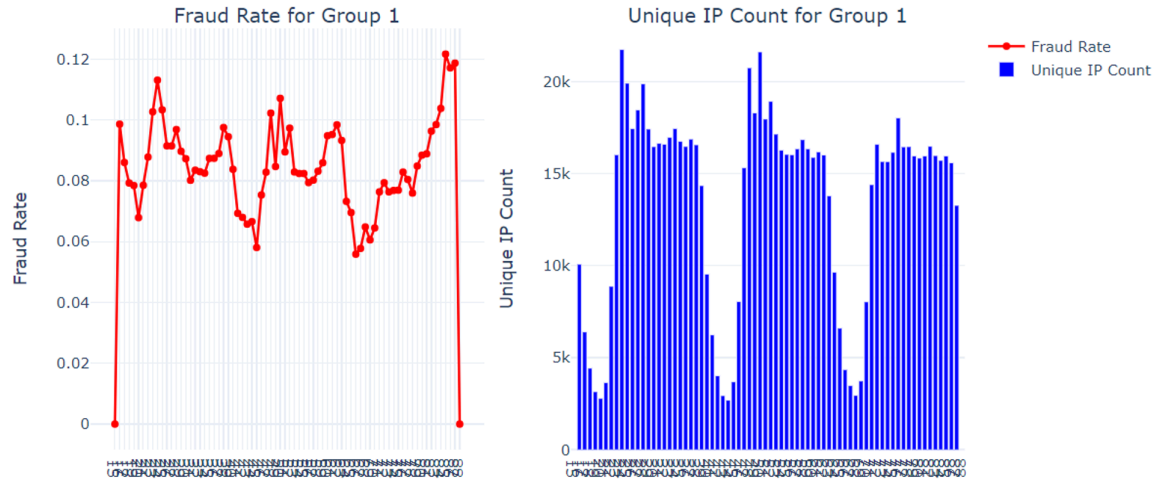

```
Detected Ratio: 1.00%, 37.51%
Detected Ratio: 2.00%, 47.70%
Detected Ratio: 3.00%, 52.97%
Detected Ratio: 4.00%, 57.64%
Detected Ratio: 5.00%, 61.65%
Detected Ratio: 6.00%, 64.12%
Detected Ratio: 7.00%, 66.70%
Detected Ratio: 8.00%, 68.21%
Detected Ratio: 9.00%, 70.29%
Detected Ratio: 10.00%, 71.49%
```

- ~~랜덤 샘플링할 때 click hour가 정렬이 돼있는지 → 속도 이슈 차 점검 필요~~

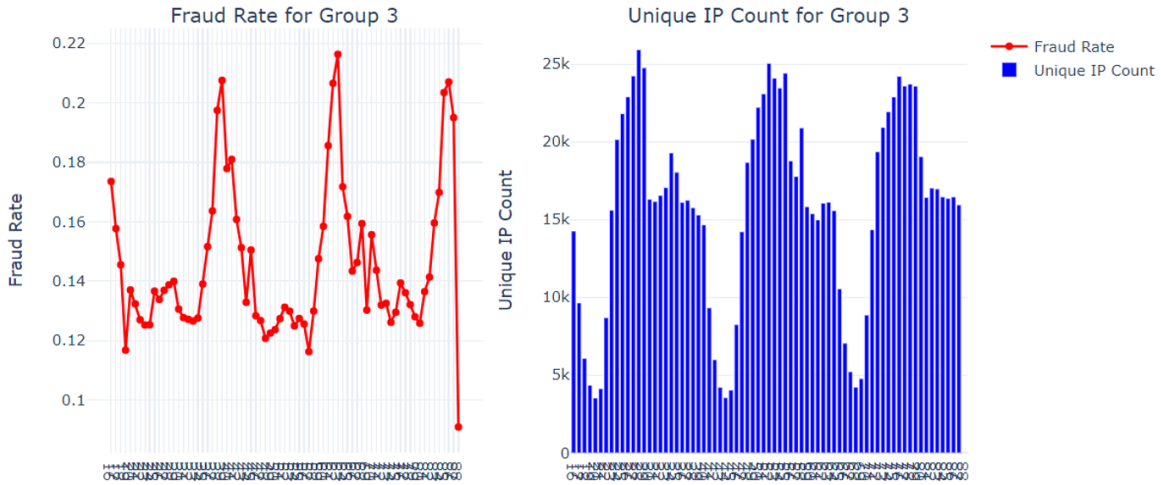
Group 0



Group 1



Group 3



~~의심: 그룹마다 왜 time_dh가 달라지는지???~~

→ **그룹별로 `time_dh` 분포가 다르지 않았음, 저번 코드가 실행이 잘못된 듯.**

app마다 time_dh

- 평균 26.96개의 `time_dh`