

K-사이버 시큐리티챌린지 2020

아이온 게임봇 유저 검출 3등 솔루션

FIND, ASK AND ANSWER

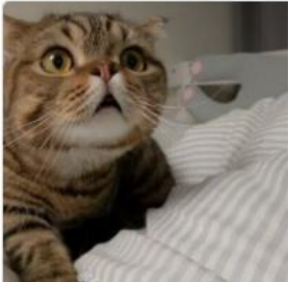
카이스트
김현우

본인 소개

- 김 현 우
- 카이스트 산업및시스템공학과 석사과정
- [TEAM-EDA 블로그](#)
- Recommender System KR 페이스북 그룹 운영진
- 깃허브 : <https://github.com/choco9966/T-academy-Recommendation>
- Kaggle : <https://www.kaggle.com/chocozzz>




본인 소개



Hyun woo kim
Student at Hanyang Univ.
Bucheon-si, Gyeonggi-do, South Korea
Joined 3 years ago · last seen in the past day
<https://eda-ai-lab.tistory.com/>

Followers 117
Following 40


Competitions
Expert

[Home](#) [Competitions \(18\)](#) [Datasets \(4\)](#) [Notebooks \(44\)](#) [Discussion \(424\)](#) [Organizations](#) [Account](#) [...](#) [Edit Profile](#)

<div>Competitions Expert</div> <div>Current Rank 1777 of 151,774</div> <div>Highest Rank 767</div> <div><div>0</div><div>6</div><div>2</div></div> <div><div>IEEE-CIS Fraud ... a year ago Top 1%</div><div>25th of 6381</div></div> <div><div>Santander Cust... 2 years ago Top 1%</div><div>38th of 8802</div></div> <div><div>Microsoft Malw... 2 years ago Top 2%</div><div>40th of 2426</div></div>	<div>Datasets Contributor</div> <div>Unranked</div> <div><div>0</div><div>0</div><div>0</div></div> <div><div>T Academy Rec... 3 months ago</div><div>5 votes</div></div> <div><div>ensemble 2 years ago</div><div>0 votes</div></div> <div><div>Gstore_v1 2 years ago</div><div>0 votes</div></div>	<div>Notebooks Expert</div> <div>Current Rank 323 of 149,384</div> <div>Highest Rank 51</div> <div><div>3</div><div>5</div><div>12</div></div> <div><div>Santander Light... 2 years ago</div><div>163 votes</div></div> <div><div>PUBG Data Des... 2 years ago</div><div>121 votes</div></div> <div><div>House Price Pre... 2 years ago</div><div>100 votes</div></div>	<div>Discussion Expert</div> <div>Current Rank 741 of 172,287</div> <div>Highest Rank 88</div> <div><div>4</div><div>3</div><div>109</div></div> <div><div>Collect all discu... 2 years ago</div><div>109 votes</div></div> <div><div>Collect discussi... 2 years ago</div><div>33 votes</div></div> <div><div>Public 19st Solu... a year ago</div><div>20 votes</div></div>
---	--	---	--

본인 소개

2020	1th Place , Prediction of the period of sale of used cars	<i>KB Capital</i>
2020	2nd Place , Winter tire Demand Forecast in Germany	<i>Hankook Tire</i>
2020	2nd Place , Detect Abnormal Transactions in Mobile Environments	<i>Pay Letter</i>
2020	2nd Place , AI Based Plant Segmentation with Aerial Photography	<i>KOFPI</i>
2020	3rd Place , Game Bot Detection in AION	<i>kisa</i>
2020	2nd Place , Korea Landmark Classification	
2020	5th Place , Nowcast Prediction	<i>Korea Hydro Nuclear Power Co.</i>
2019	4th Place , Fire Prediction in Gimhae	<i>LH Co.</i>
2019	3rd Place , Prediction Jeju Bus Passengers	<i>Jeju Technopark</i>
2019	5th Place , Automotive network intrusion detection	<i>Kisa</i>
2019	4th Place , Brunch Recommendation Competition	<i>Kakao</i>
2019	1st Place , Prediction the real price of apartments	<i>Zigbang</i>
2018	1st Place , Data Visualization Challenge of Credit Card transaction	<i>Banksalad</i>
2018	1st Place , Bigcontest2018	<i>Shinhan Bank</i>

본인 소개

2020	1th Place , Prediction of the period of sale of used cars	KB Capital
2020	2nd Place , Winter tire Demand Forecast in Germany	Hankook Tire
2020	2nd Place , Detect Abnormal Transactions in Mobile Environments	Pay Letter
2020	2nd Place , AI Based Plant Segmentation with Aerial Photography	KOFPI
2020	3rd Place , Game Bot Detection in AION	kisa
2020	2nd Place , Korea Landmark Classification	
2020	5th Place , Nowcast Prediction	Korea Hydro Nuclear Power Co.
2019	4th Place , Fire Prediction in Gimhae	LH Co.
2019	3rd Place , Prediction Jeju Bus Passengers	Jeju Technopark
2019	5th Place , Automotive network intrusion detection	Kisa
2019	4th Place , Brunch Recommendation Competition	Kakao
2019	1st Place , Prediction the real price of apartments	Zigbang
2018	1st Place , Data Visualization Challenge of Credit Card transaction	Banksalad
2018	1st Place , Bigcontest2018	Shinhan Bank

INDEX

ANALYSIS

1. 탐색적 데이터 분석
2. 기존 논문 리서치



MODEL

1. 파생변수 생성
2. 파생변수에 대한 정규화
3. 학습 전략 및 검증방법 생성
4. 모델링

RESULTS

1. 변수중요도에 대한 해석
2. 성능에 대한 분석
3. 모델의 장단점과 한계



PROBLEM DESCRIPTION

게임 데이터를 분석하여 높은 정확도로 게임봇을 탐지할 수 있는 머신러닝 & AI 기반 알고리즘 개발¹

- MMORPG와 온라인 게임 내 재화, 아이템의 환금성을 악용하여 대량의 캐릭터를 운용하여 수입을 얻는 "작업장"은 게임 밸런스를 해치고 이용자에게 불편을 야기함
- 작업장에서 이용하는 "게임봇"은 환금성이 있는 재화나 아이템을 채굴하기 위해 특정 행위를 반복

게임봇(Game Bot)² : 사람을 대신하여 자동으로 게임플레이를 해주는 프로그램

- 게임 내 설계된 콘텐츠를 빠르게 소진하여 게임 수명이 짧아지게 만드는 요인 중 하나
- 일반 유저의 게임 플레이에 방해가 되어 불평 발생



news.mt.co.kr > mtview ▼

"핵 때문에 못해먹겠다"...배그 떠나는 유저들 - 머니투데이

2020. 10. 3. — "배틀그라운드(배그)는 핵만 아니었으면 피파온라인4는 물론 롤(리그오브레전드) ... 핵 사용자들은 조준을 하지 않아도 상대를 100% 명중시킬 수 있는 등 ... 직접 피해액은 매출액 감소에 따른 게임사들의 피해 1조1921억원, 게임 핵 ...

www.insight.co.kr > news ▼

핵 사용자 너무 많아 매출 폭망+접속자 반토막 나버린 '배그 ...

2019. 11. 20. — 배틀그라운드가 게임 내 핵 사용자들로 인해 곤욕을 치르고 있다. ... 매출은 예전만 못하고 접속자 수는 나날이 감소하고 있다. ... 최근 한 10대 학생이 배그에서 사용 가능한 변종 '게임 핵'을 만들어 유저들에게 판매해 벌금형을 선고 ...

1) 고려대학교, 정보보안학과 해킹탐지 연구소 (2020)
2) 데이터분석 기반 게임봇과 작업장 탐지 (NDC 2017)

DATA DESCRIPTION

아이온 게임 유저 액션로그와 유저별 정상/게임봇 여부를 제공

- 예선 : 10일간의 게임로그와 분석용/제출용 유저 목록 (2010.05.08 ~ 2010.05.17) – 약, 70GB
- 본선 : 7일간의 게임로그와 제출용 유저 목록 (2020.05.18 ~ 2020.05.24)

log_date	big_log_id	log_id	actor	actor_account	target	target_account	worldnum	location_x	location_y	...	etc_num7	etc_num8
2010-05-08 00:00:00.093	100	160	149008	1051624	0	0	2147494864	229	304	...	3514.0	146070032.0
2010-05-08 00:00:03.750	100	103	109510	1043965	0	0	220050000	1303	1950	...	803208.0	0.0

1번째 Row에 대한 설명

- log_id : 캐릭터정보
- actor : 캐릭터 UID
- actor_account : 계정 UID

·
·
·

- etc_num7 : DP
- etc_num8 : EXP

2번째 Row에 대한 설명

- log_id : 아이온 게임에 접속
- actor : 캐릭터 UID
- actor_account : 계정 UID

·
·
·

- etc_num7 : 활력 포인트
- etc_num8 : 구원의 기운 적용 여부

ANALYSIS



1. 탐색적 데이터 분석 (Exploratory Data Analysis)

1.1. 일반유저와 게임봇의 활동 시간 비교

1.2. 일반유저와 게임봇의 플레이 스타일 비교

2. 기존 문헌 분석

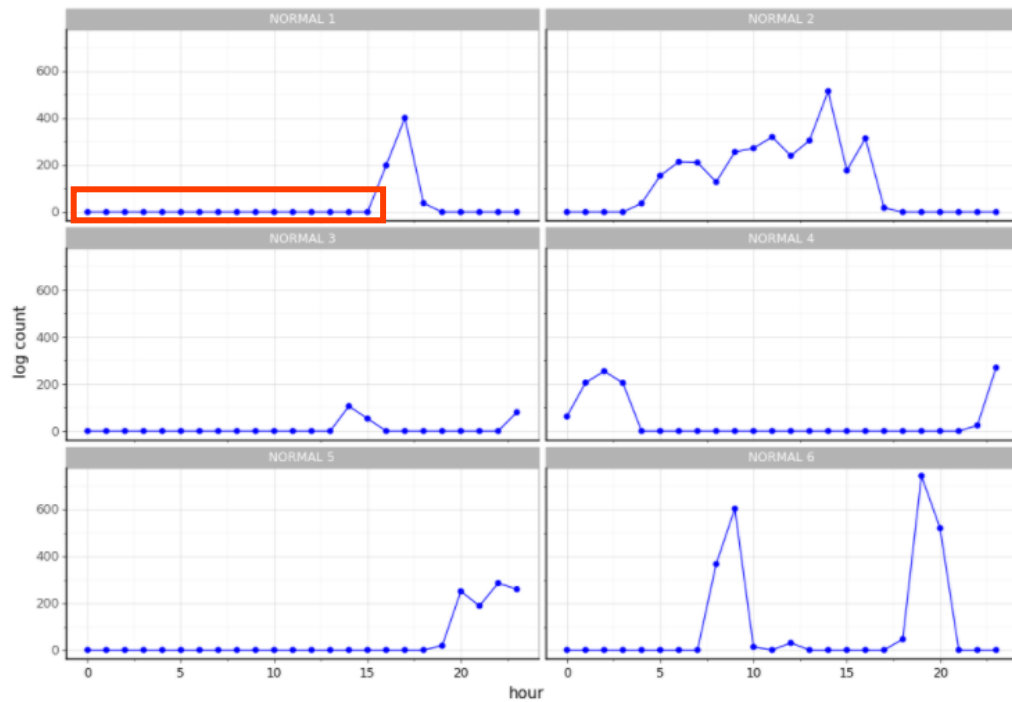
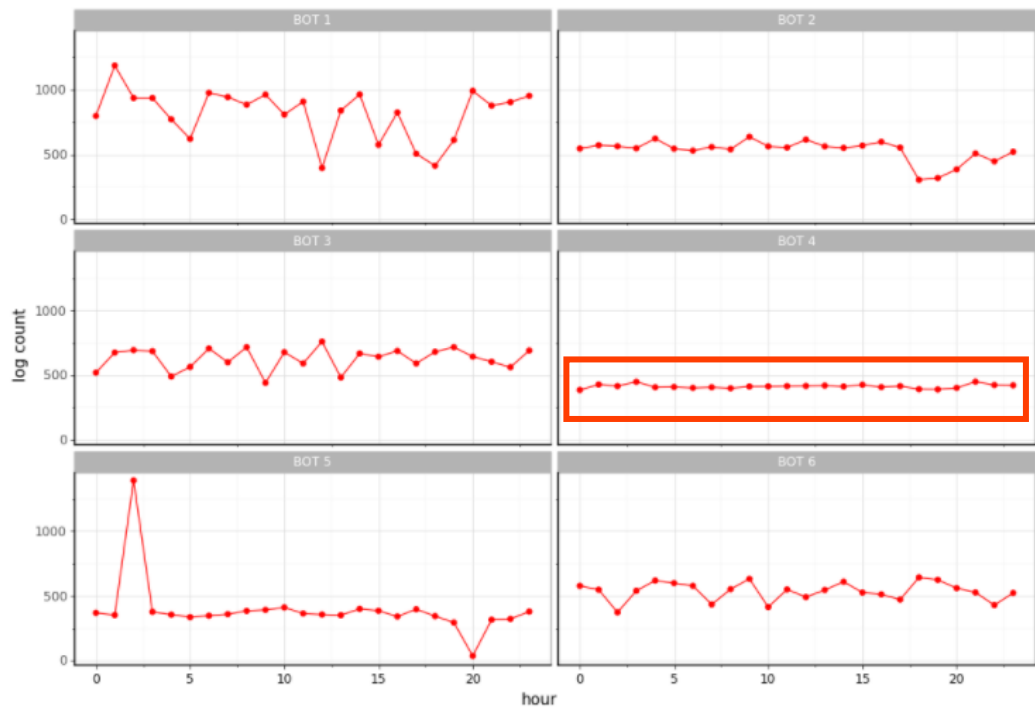
2.1. Multimodal game bot detection using user behavioral characteristics(2016)

2.2. 자산변동 좌표 클러스터링 기반 게임봇 탐지(2015)

2.3. 자기 유사도를 이용한 MMORPG 게임봇 탐지 시스템(2016)

1. 탐색적 데이터 분석

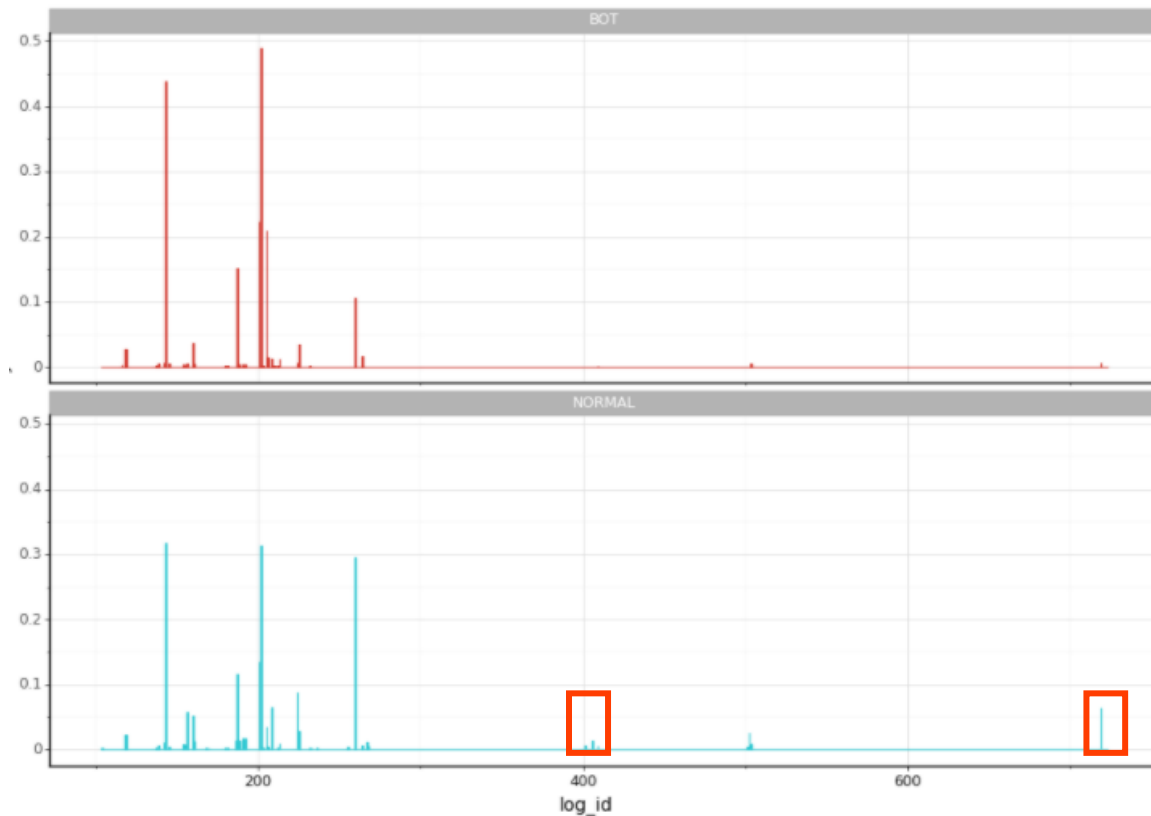
• (1) 일반유저와 게임봇의 활동 시간 비교



- 게임봇은 매시간 플레이를 하고 몇몇 게임봇은 시간당 로그의 숫자가 거의 비슷함
- 일반유저는 매시간 플레이를 하지 않고, 로그의 숫자가 0이 되는 종료시간이 존재
 - 쉬는 시간과 플레이시간이 구분 되어있음

1. 탐색적 데이터 분석

• (2) 일반유저와 게임봇의 플레이 스타일 비교 – 상위 5개의 로그



게임봇

- 아이템 획득 (25%)
- 경험치 획득 (23%)
- 아이템 생성 (11%)
- 아이템 채집 (11%)
- 키나 증가 (8%)

일반유저

- 경험치 획득 (16%)
- 아이템 획득 (16%)
- 아이템 삭제 (15%)
- 아이템 생성 (7%)
- 키나 증가 (6%)

일반유저

- 다양한 로그의 플레이 활동을 진행
- 400번대 (스킬 관련)와 700번대 (공성전 관련)의 플레이 활동을 진행

게임봇

- 사냥과 관련된 로그에 집중 되어있음
- 채집과 같이 귀찮은 활동을 더 많이 함

2. 탐색적 데이터 분석



(1) Multimodal game bot detection using user behavioral characteristics

Table 2 Personal and social features

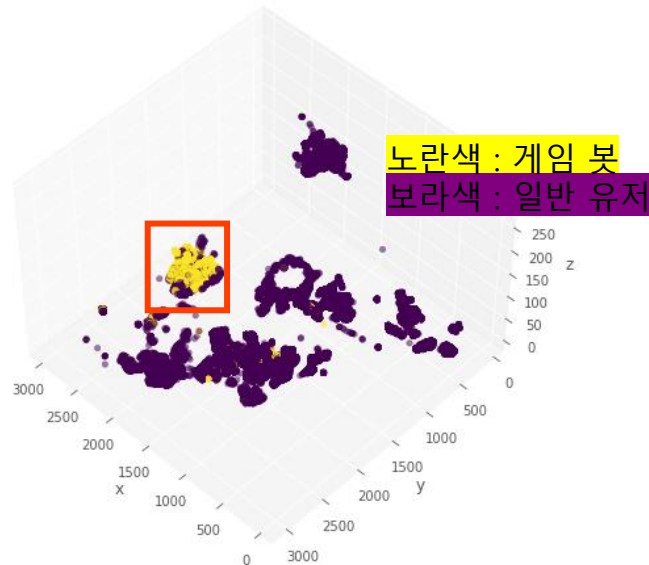
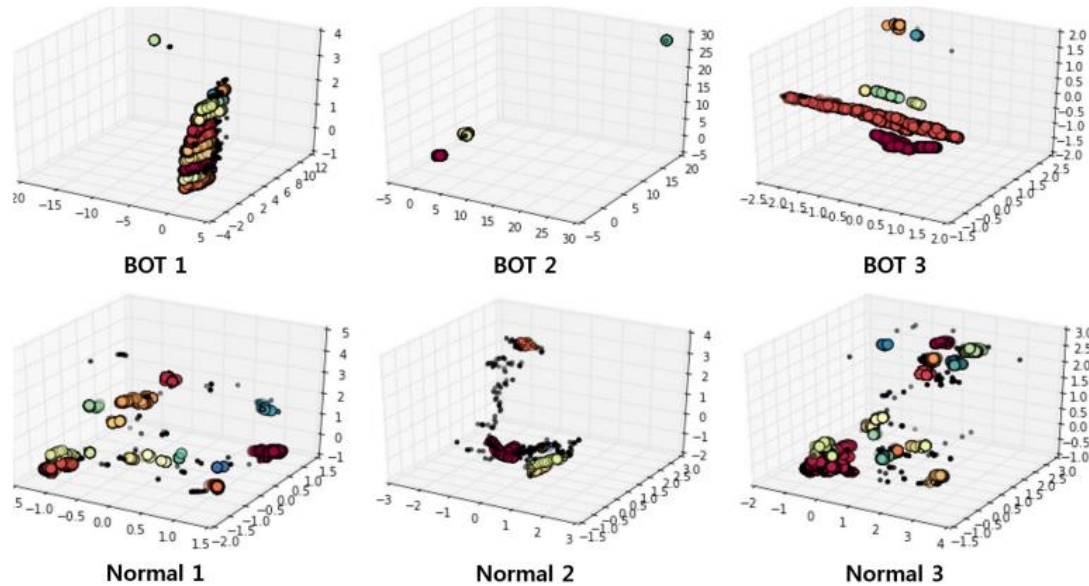
Category	Key idea
<i>Personal feature</i>	
Player information	Login frequency, play time, game money, number of IP address
Player actions	Sitting, earning experience points, obtaining items, earning game money, earning player kill points, harvesting items, resurrection, restoring experience points, being killed by a non-player and/or player character, using portals
<i>Social feature</i>	
Group activities	Party play time, guild activities
Social interaction diversity	Party play, friendship, trade, whisper, mail, shop, guild
Network measures	Degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, eccentricity, authority, hub, PageRank, clustering coefficient

플레이어의 행동을 4가지 유형으로 나누어서 파생변수를 생성

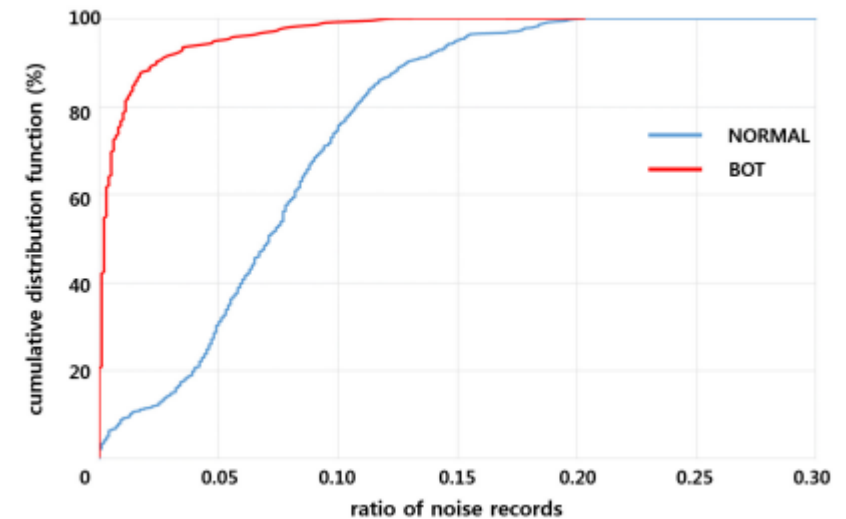
1. Player Information
2. Player Action
3. Group Activities
4. Network Measures

2. 탐색적 데이터 분석

(2) 자산변동 좌표 클러스터링 기반 게임봇 탐지



- 게임봇은 일반 유저에 비해서 활동반경의 범위가 좁고 특정 공간내에서만 반복해서 움직이는 특징을 보임
- 이러한 좌표를 모델에 반영해 주기위해서 소지금 변동 위치 좌표의 공간적 특징을 DBSCAN 알고리즘을 통해서 추출



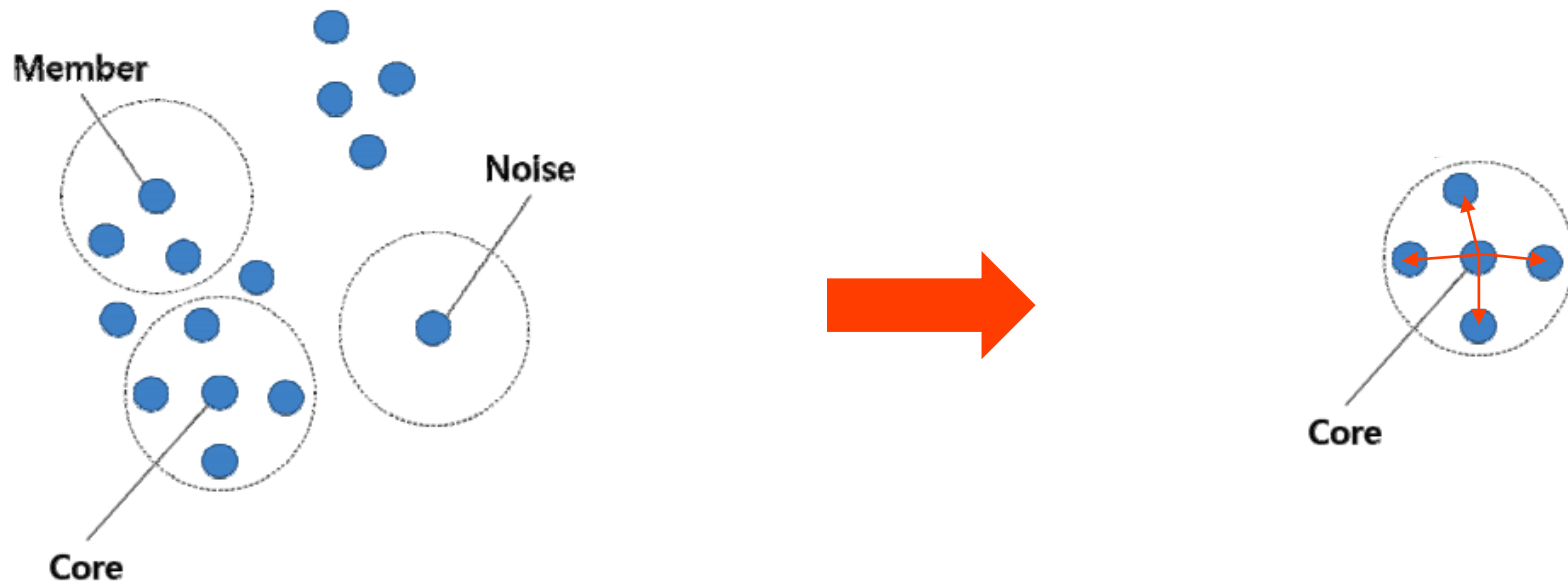
2. 탐색적 데이터 분석



2. 자산변동 좌표 클러스터링 기반 게임봇 탐지

원문 논문 : DB SCAN을 이용해서 CORE, MEMBER, NOISE를 파생변수로 활용

- 속도가 오래 걸리는 단점이 존재



수정된 파생변수 : 중심 좌표를 기반으로 거리의 표준편차, 평균을 계산

- 기존 대비 10배의 속도로 개선
- 성능도 기존 대비 큰 차이가 없음

2. 탐색적 데이터 분석



(3) 자기 유사도를 이용한 MMORPG 게임봇 탐지 시스템

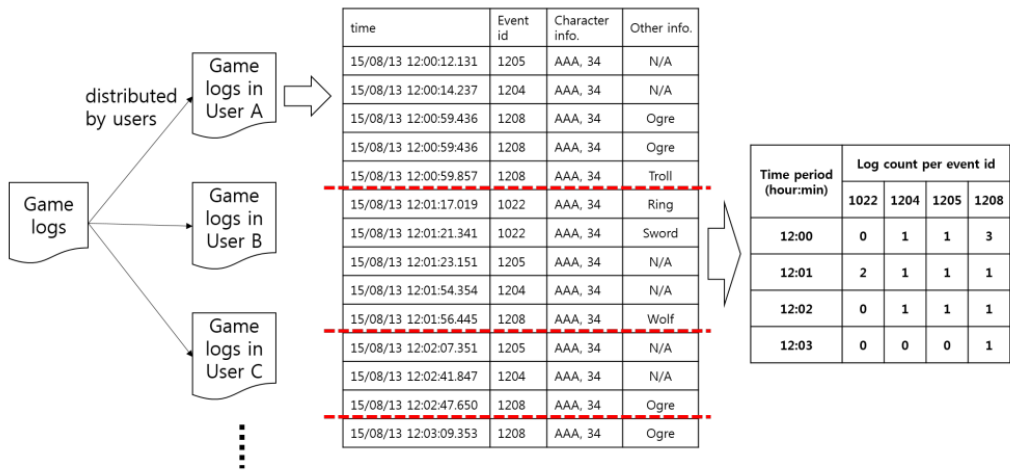


Fig. 11. Process for transforming game logs to vectors. In the above, four vectors are generated: (0, 1, 1, 3), (2, 1, 1, 1), (0, 1, 1, 1), (0, 0, 0, 1).

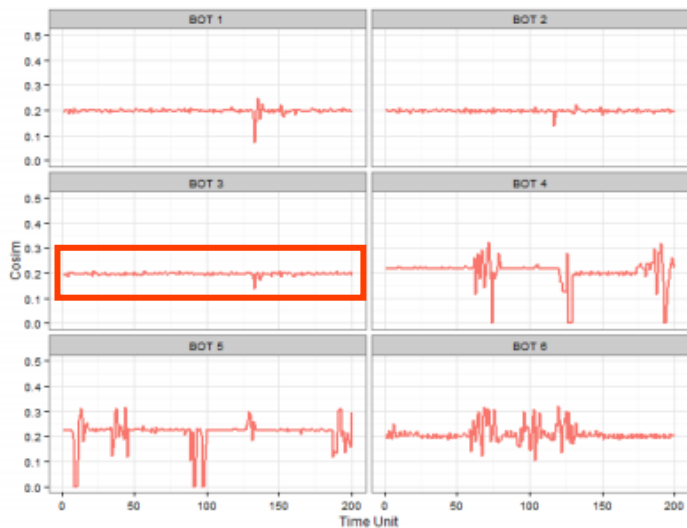
Time Period	1번 로그	2번 로그	...	N번 로그
10/05/08 12:00	1	4		8
10/05/08 12:01	1	4		8
⋮				
15/05/17 23:09	1	4		8

- 게임봇은 일반 유저에 비해서 같은 로그들을 반복하는 행동을 보임
- 분단위로 유저가 행동한 로그를 기록
- 단위 벡터와 코사인 유사도를 모든 시간에 대해서 계산 (자기유사도)

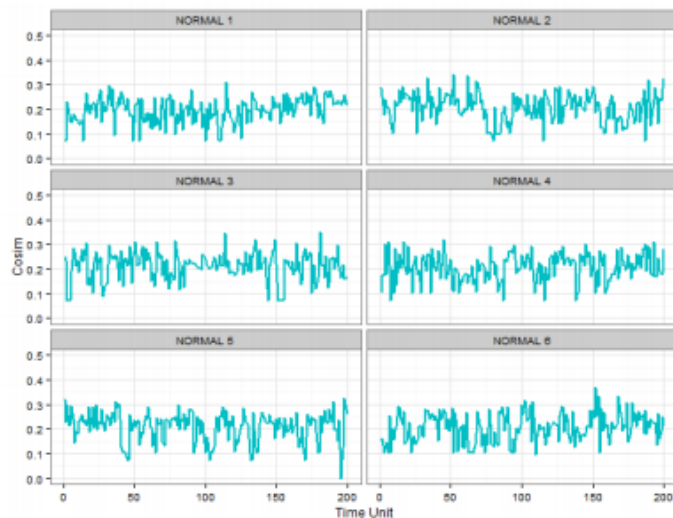
- 모든 로그에 대한 단위 벡터 : $[1, 1, \dots, 1]$ 를 준비
- 단위 벡터와 코사인 유사도를 계산
 - 자기 유사도 : $[\text{코사인 유사도}_1, \dots, \text{코사인 유사도}_M]$

2. 탐색적 데이터 분석

• (3) 자기 유사도를 이용한 MMORPG 게임봇 탐지 시스템



(a) bots



(b) normal users

- 게임봇은 자기유사도가 분단위로 굉장히 비슷한 모습을 보임
- 일반 유저는 자기유사도가 굉장히 다름
- 이를 이용해서 모든 분마다의 자기유사도의 표준편차를 계산

MODEL



1. EDA와 리서치를 통한 파생변수 생성
2. 정규화를 통한 Covariate Shift 방지
3. 학습전략 및 검증방법 생성
4. LightGBM을 이용한 모델링

2. MODEL



• (1) EDA와 리서치를 통한 파생변수 생성

Player Information

- 접속 로그가 있는 날의 수, 일별 로그인 횟수, 로그아웃 횟수 , IP의 개수, 플레이 시간, 최고 레벨 , 평균 레벨
-

Player Action

- 앓은 횟수, 경험치 획득량, 횟수(로그의 수), 아이템 획득량, 채집량, 채집횟수, 채집간격, 경험치 복구량, 복구 횟수
 - 돈 획득량, 포털 이용 횟수, 텔레포트 사용 횟수, PC/NPC에게 죽은 횟수, 퀘스트 완료 횟수
 - 솔로 사냥으로 얻은 키나의 수, 파티 사냥으로 얻은 키나의 수, 아이템 삭제의 수
-

Group Activity

- 총 파티 시간
-

Social Activity

- 거래 신청 보낸 횟수, 거래 신청 수락 횟수, 상점에서 구입한 횟수, 상점에서 판매한 횟수
 - 개인상점에서 구매한 횟수, 개인상점에서 판매한 횟수, 우편을 보낸 횟수, 우편을 받은 횟수
 - 파티 초대 신청을 보낸 횟수, 파티 초대 신청을 받은 횟수, 결투 횟수, 친구 수
 - 시간당 획득량이 과도한 경우, 키나를 보낸 총 금액, 키나를 받은 총 금액
-

Research

- 자기유사도 기반의 파생변수, 유니크한 분단위 로그의 수, 자산변동 좌표 클러스터링 기반의 거리의 표준편차

2. MODEL



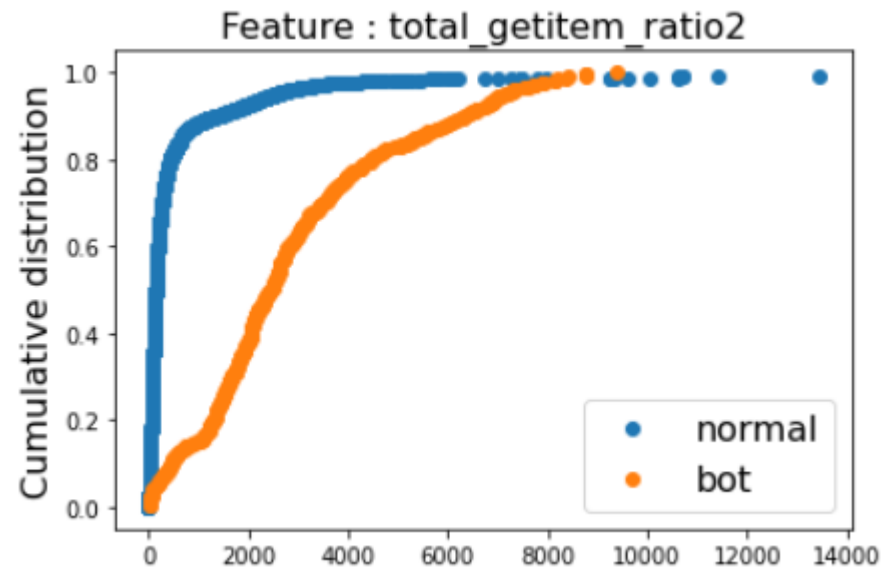
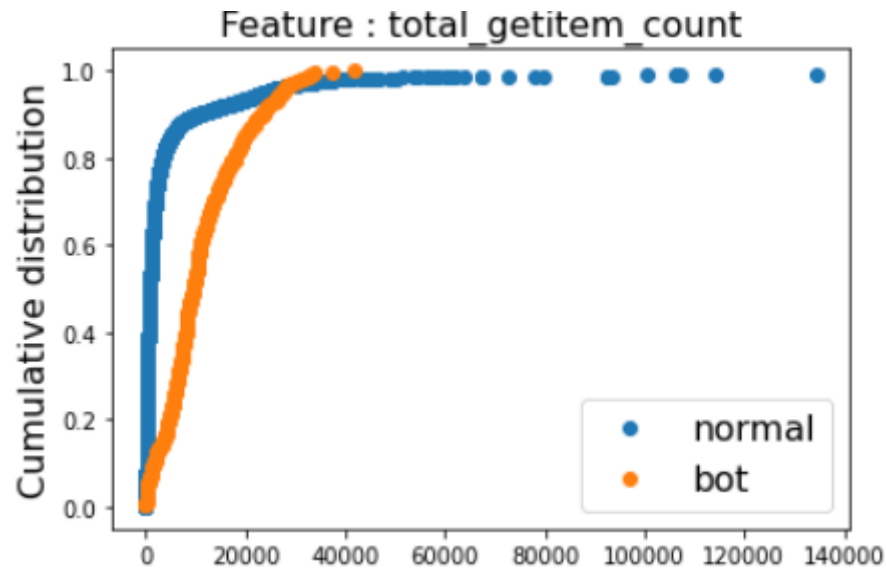
(2) 정규화를 통한 Covariate Shift 방지

논문¹의 방법을 참고하여, 일별, 주별 행위의 Aggregation 사용

- 단순히 SUM, COUNT 기반의 파생변수는 아래의 한계점이 존재
- 데이터의 모수가 다른 경우에 분포의 차이를 보임 (17일 데이터에서 1일 플레이 vs 17일 플레이)
- 이를 해결하기 위해서, Ratio기반의 보조 변수를 사용

Ratio 기반의 Feature Engineering

- Ratio1 : 전체 로그의 수로 나눈 경우
- Ratio2 : 전체 플레이 시간으로 나눈 경우
- Ratio3 : 전체 플레이 날의 수로 나눈 경우



2. MODEL

(3) 학습전략 및 검증방법 생성

Class Imbalance – 게임봇의 유저가 너무 적음

```
1 labeled = pd.read_csv("./Dataset/labeled_account.csv")  
2 labeled['class'].value_counts(normalize=True)
```

```
0    0.925926  
1    0.074074  
Name: class, dtype: float64
```

샘플링

전부사용

학습데이터셋1

일반유저

게임봇

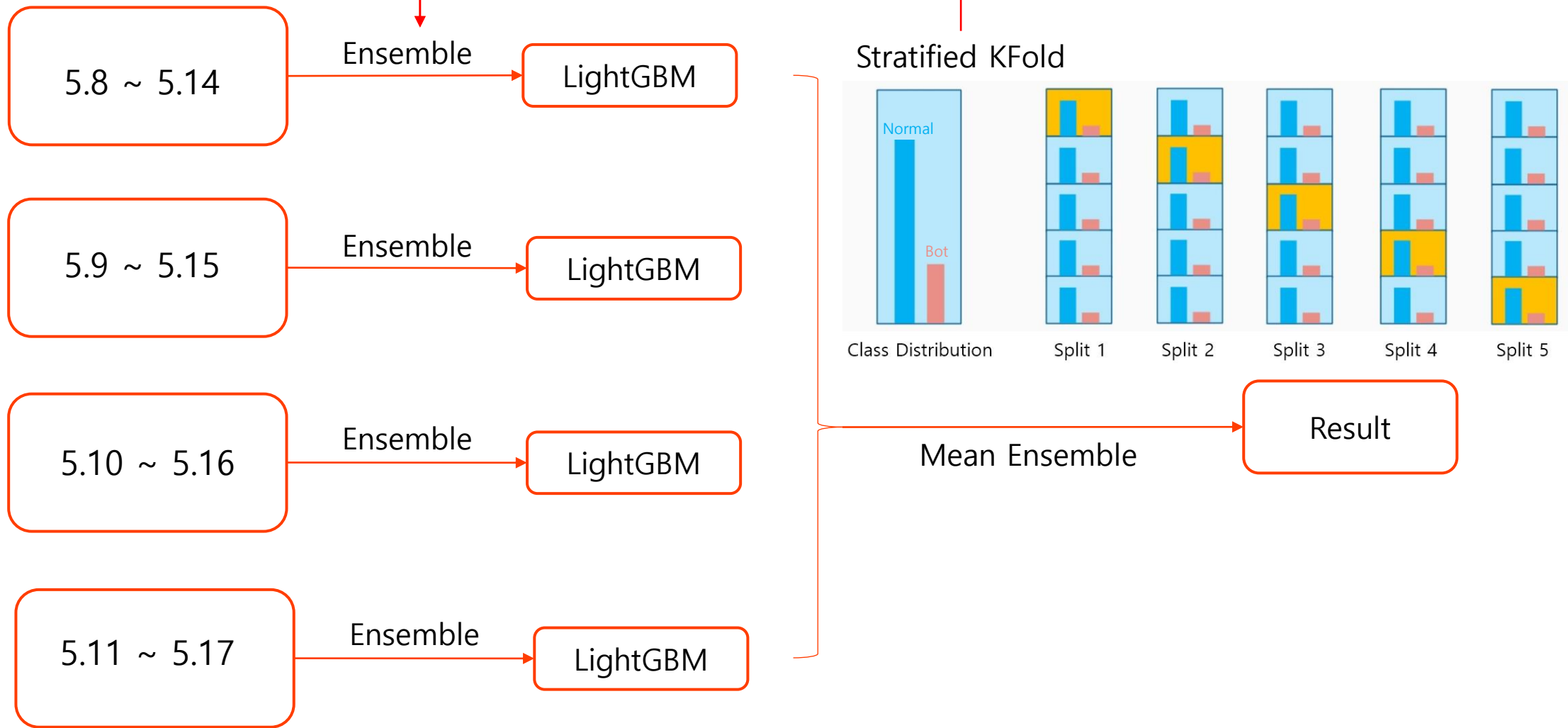
학습데이터셋N

...

2. MODEL

(4) LightGBM을 이용한 모델링

Unbalanced를 반영하기 위해서, 게임봇이 아닌 유저는 샘플링해서 사용
(신뢰성을 보장하기 위해 여러 번 샘플링하고 평균)



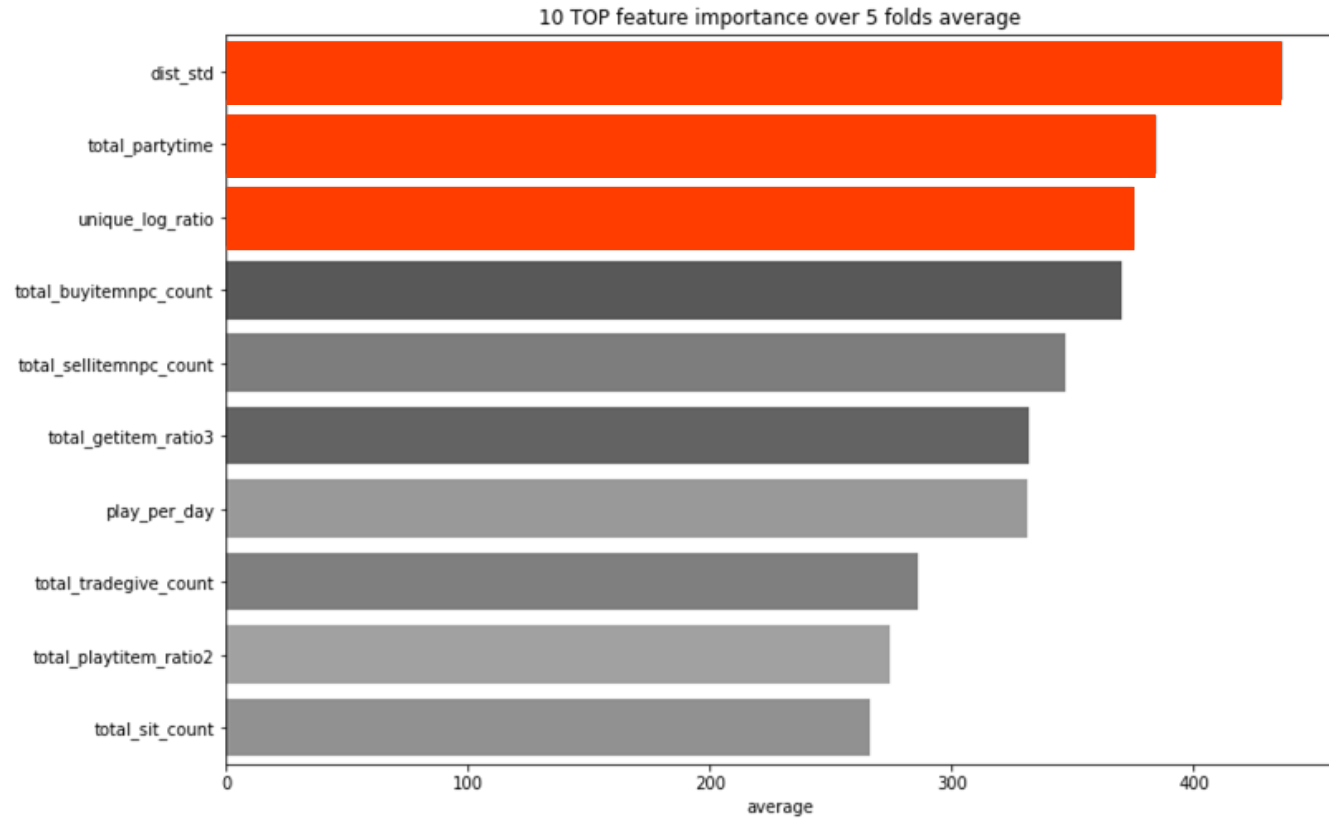
RESULTS



1. 변수 중요도에 대한 해석
2. 모델 성능에 대한 분석
3. 모델의 장단점과 한계

1. 변수 중요도에 대한 해석

• (1) LightGBM 중요도

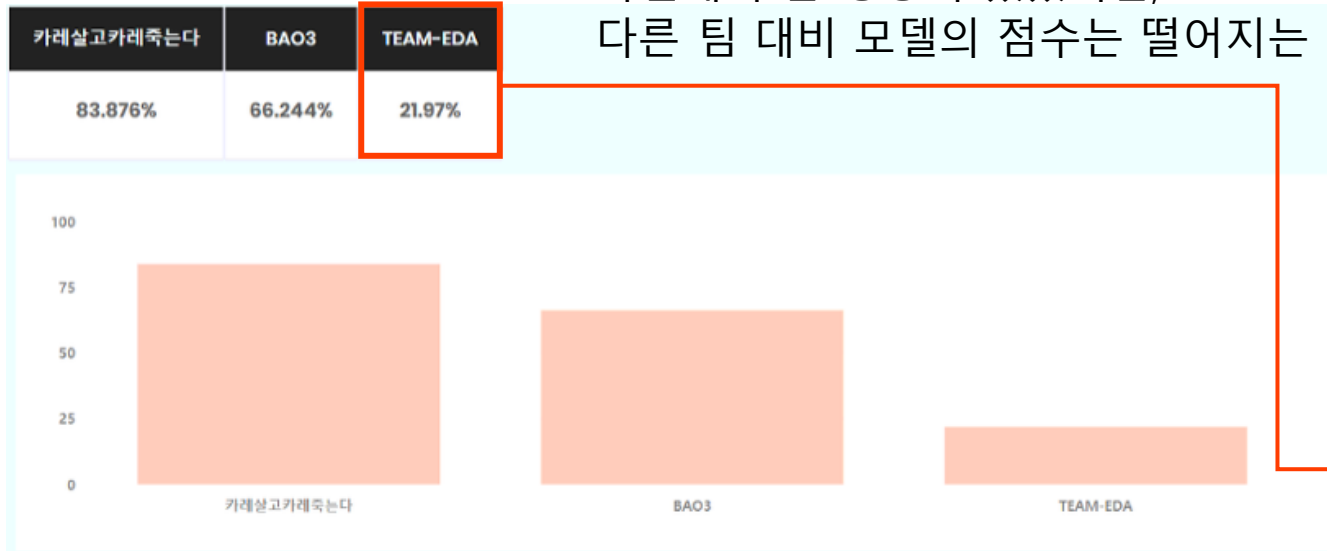


- Player Information, Action, Group Activities, Social 중에서 Player 관련 정보가 가장 중요도가 높고 Research에서 만든 변수들은 중요도가 전부 상위권
- Ratio 기반의 파생변수도 중요도에서 높은 것을 확인할 수 있음

2. 성능에 대한 분석

(1) 최종 결과

작년대비 큰 성장이 있었지만,
다른 팀 대비 모델의 점수는 떨어지는 편

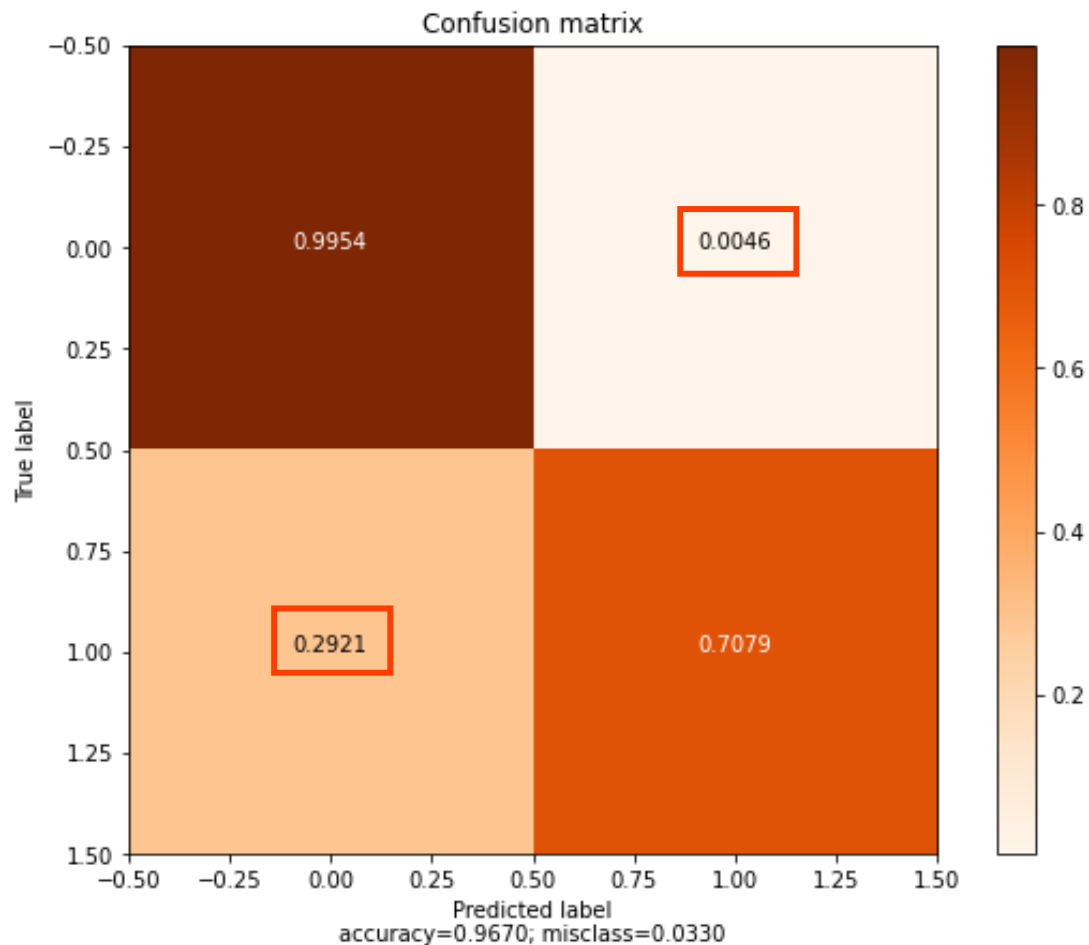


게임봇 탐지		
순위	팀 이름	총 점수
1	아이씨유	87.50%
2	밥2조	84.20%
3	Smiley Mk.3	83.90%
4	에스엔티웍스	82.80%
5	TEAM-EDA	78.60%

2. 성능에 대한 분석

(2) Precision 과 Recall

운영상에서는 게임봇을 맞추는 것도 중요하지만, 일반 유저를 게임봇으로 분류하지 않는 것도 중요 !!



- 보수적으로 게임봇과 일반유저를 분류하도록 모델을 설계
 - 게임봇 -> 게임봇 : 70.79%
 - 게임봇 -> 일반 유저 : 29.21%
 - 일반유저 -> 게임봇 : 0.46%**
 - 일반유저 -> 일반유저 : 99.54%
- 점수는 다른 팀 대비해서 떨어지지만, 일반유저를 게임봇으로 분류하는 실수는 거의 없음

3. 모델의 장단점과 한계

(1) 모델의 장단점

장점

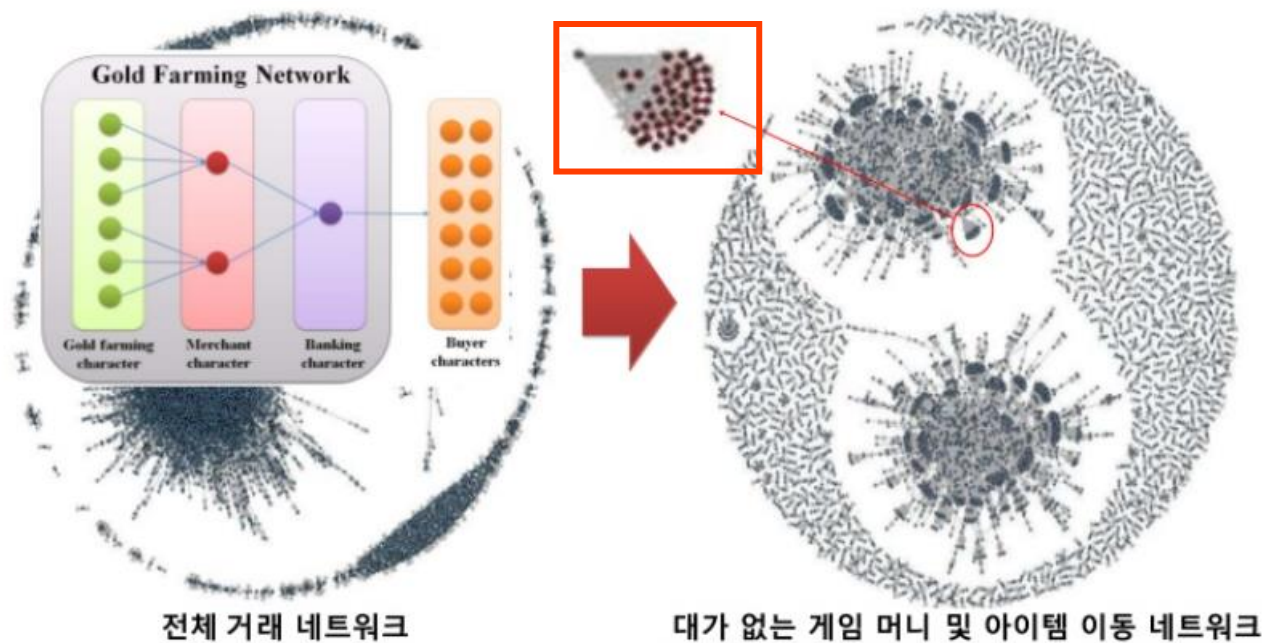
- 사람을 게임봇으로 분류하는 비율이 적기때문에, 운영상에서 유리하다.
- 주단위로 모델을 생성하기에, 시간이 흘러 게임봇의 특징이 달라져도 모델이 학습할 수 있다. (유지보수에 강함)

단점

- 다른 팀들에 비해서 모델의 성능이 낮다. (게임봇을 탐지하지 못한다.)
- 기존에 제안한 방법들을 기반으로 모델을 생성해서 노벨티가 떨어진다.
- 여러 개의 모델을 생성해야 하므로 자원측면에서 아쉬움이 존재한다.
- 매일 혹은 매주 모델을 학습해야 한다.

3. 모델의 장단점과 한계

(2) 시도한 방법들과 한계 - 네트워크 기반의 파생변수



- 1, 2에서 네트워크 중요도 언급 1에서 특히 대가성 없는 게임 머니 이동 언급
- 이를 잘 잡기위해서 2에서 Network 기반의 Node edge degree eigen value 등등의 피쳐 제안
- 하지만 결과적으로 성능은 떨어짐

Table 4 Basic network characteristics of six interaction networks

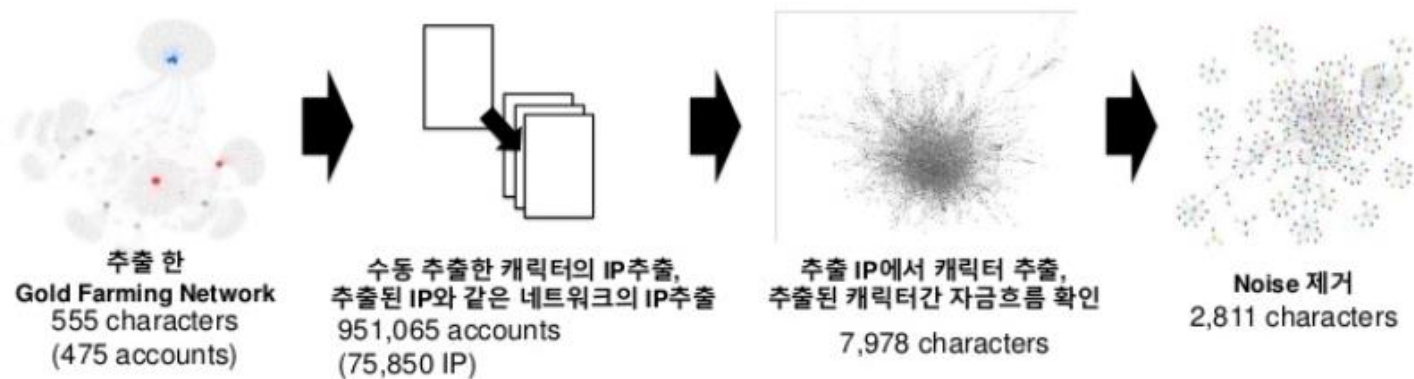
	Party		Friendship		Trade		Whisper		Mail		Shop	
	Bot	Human	Bot	Human	Bot	Human	Bot	Human	Bot	Human	Bot	Human
Nodes	1756	33,924	479	24,628	4003	30,640	434	16,209	4848	28,362	305	7001
Edges	2463	862,021	749	174,626	9809	162,236	656	248,133	12,873	76,844	362	11,824
Avg. degree	1.4	25.41	1.56	7.09	2.45	5.29	1.51	15.31	2.66	2.71	1.19	1.7
Network diam.	22	15	9	15	25	18	23	12	9	24	5	28
Avg. C.C.	0.1	0.07	0.07	0.09	0.41	0.08	0.01	0.05	0.12	0.19	0.12	0.01
Avg. path len.	6.14	3.77	2.18	4.7	5.66	5.41	6.41	3.65	2.16	7.55	1.58	8.14

The average degree of all interaction networks of the human group is higher than that of the game bot group. This shows that game bots do not enjoy socializing with other users

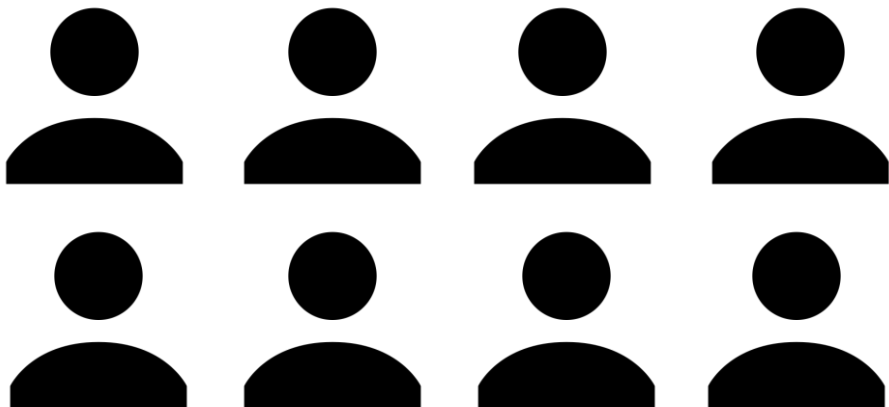
1) 데이터분석 기반 게임봇과 작업장 탐지 (NDC 2017)
 2) Multimodal game bot detection using user behavioral characteristics

3. 모델의 장단점과 한계

(2) 시도한 방법들과 한계 - IP 기반의 게임봇 탐지



게임봇 유저들의 IP 목록



게임봇 여부를 모르는 유저의 IP목록



IP : 192.168.xxx.xxx

.

.

.

IP : 192.168.xxx.xxx

추출한 IP가 게임봇 유저들의 IP 목록에 얼마나 겹치는 지

3. 모델의 장단점과 한계

(2) 시도한 방법들과 한계 - 모델 및 방법론



모델

- 부스팅 모델 : XGBOOST, CATBOOST
- 배깅 모델 : RANDOM FOREST
- 선형 모델 : LOGISTIC REGRESSION
- 딥러닝 모델 : LSTM, DNN
- 커널 모델 : SVM

방법론

- PSEUDO LABELING
- SEED ENSEMBLE
- STACKING
- WEEK차별로 결과의 MAX

참고 문헌

- Kang, A. R., Jeong, S. H., Mohaisen, A., & Kim, H. K. (2016). Multimodal game bot detection using user behavioral characteristics. *SpringerPlus*, 5(1), 1-19.
- Show me your Account Detecting MMORPG Game Bot Leveraging Financial Analysis with LSTM
- Multimodal game bot detection using user behavioral characteristics
- 온라인 게임 내의 부정 행위 탐지 연구 동향
- 자기 유사도를 이용한 MMORPG 게임 봇 탐지 시스템
- 자산변동 좌표 클러스터링 기반 게임 봇 탐지
- 행위 시간 간격 기반 게임 봇 탐지 기법
- Permutation importance: a corrected feature importance measure
- 2019년도 게임봇 1등 솔루션
- 2019년도 게임봇 2등 솔루션
- 2019년도 게임봇 3등 솔루션

감사합니다