Categorical Time Series Prediction with Embeddings

김지오. 박준우. 이승재. 장윤태. 조수연. 조영규

목차

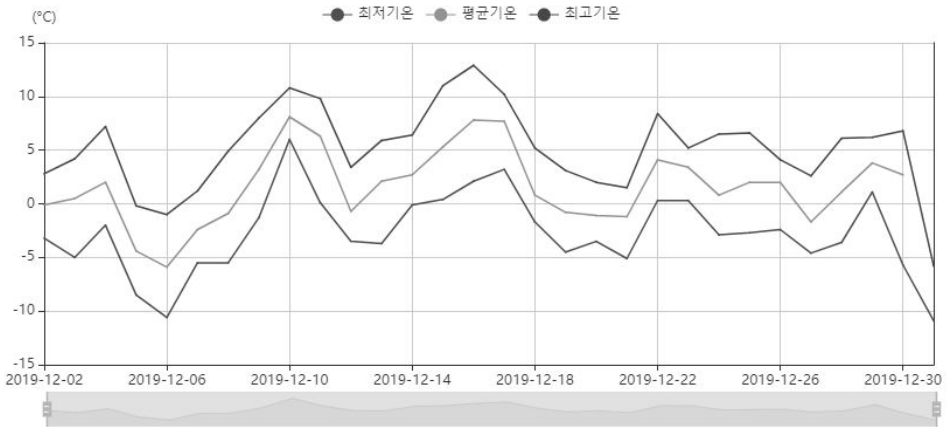
1

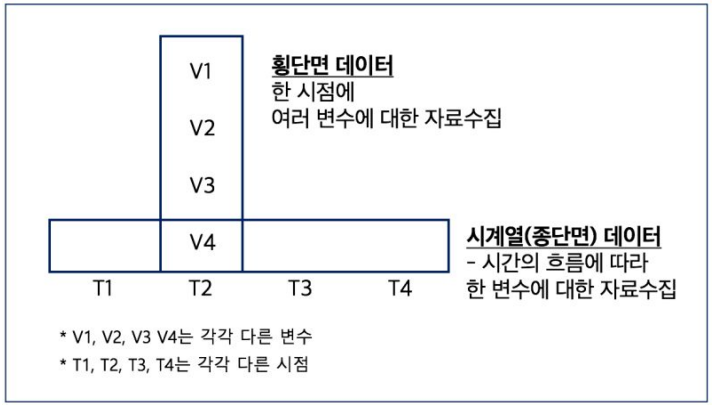
모델링 목적

Project Objective

시계열이란?

시간의 흐름에 따라 일정한 간격 (년도. 분기. 월. 일별 등)으로 기록된 (종단면) 데이터





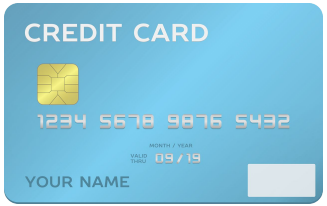
https://m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=her7845&logNo=220878719799

https://domini21.tistory.com/14

모델링 주제

제품 검색

장바구니 추가



제품 결제

상품 목록 확인

Next Move?

플랫폼 내에서의 고객 행동을 기반으로 다음 행동을 예측하여 Personalized한 서비스 제공할 수 있는 기회 제공

2

데이터셋 & 모델 구조

Dataset & Model Architecture

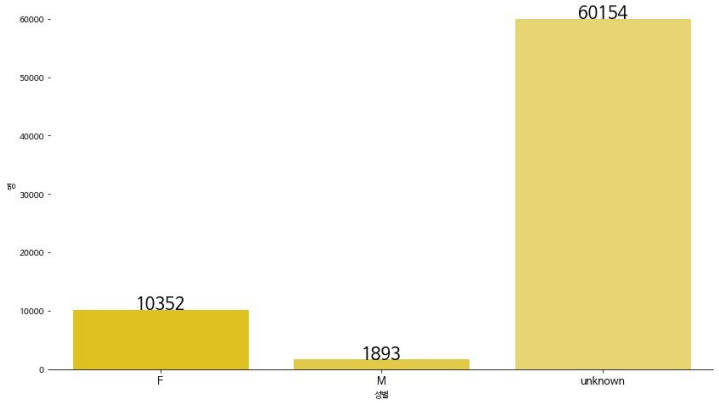
데이터셋

L.POINT 고객 Demo 정보 데이터 설명 테이블

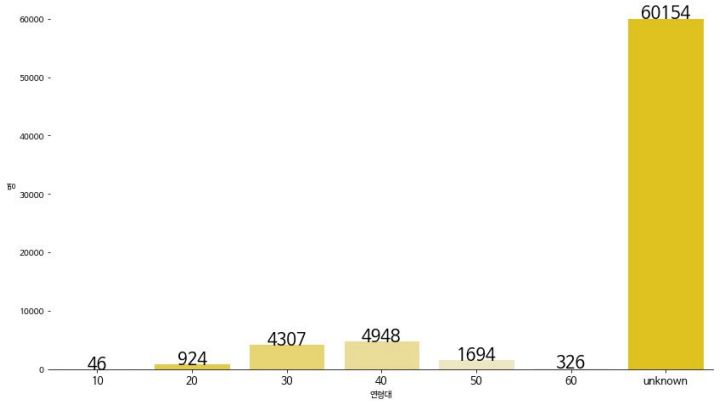
No	변수명(영문)	변수명(국문)	상세설명	PK
1	CLNT_ID	클라이언트ID	고객을 고유하게 식별할 수 있도록 랜덤으로 부여된 ID	Y
2	CLNT_GENDER	성별	성별정보 [남자: M/ 여자: F/ 정보없음 : unknown]	
3	CLNT_AGE	연령대	연령대 정보 [10대이하/ 20대 / 30대 / 40대 / 50대 / 60대이상 / 정보없음: unknown]	

Row: 72399
→ 고객 72399명의 Demo

고객 Demo 데이터 성별 분포



고객 Demo 데이터 연령 분포



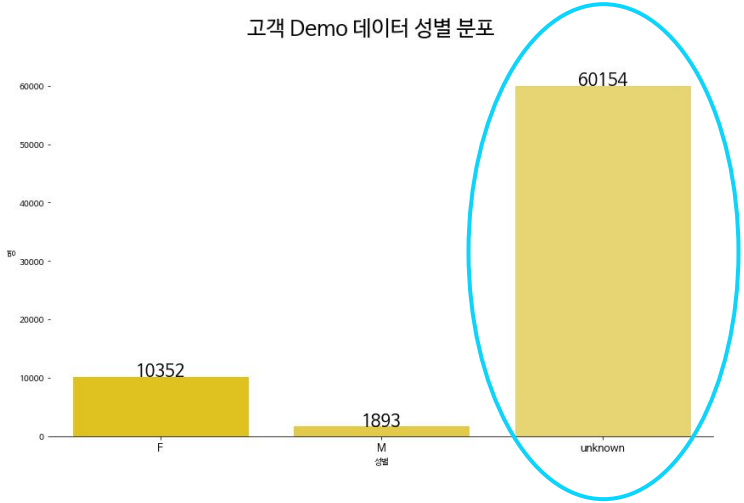
데이터셋

L.POINT 고객 Demo 정보 데이터 설명 테이블

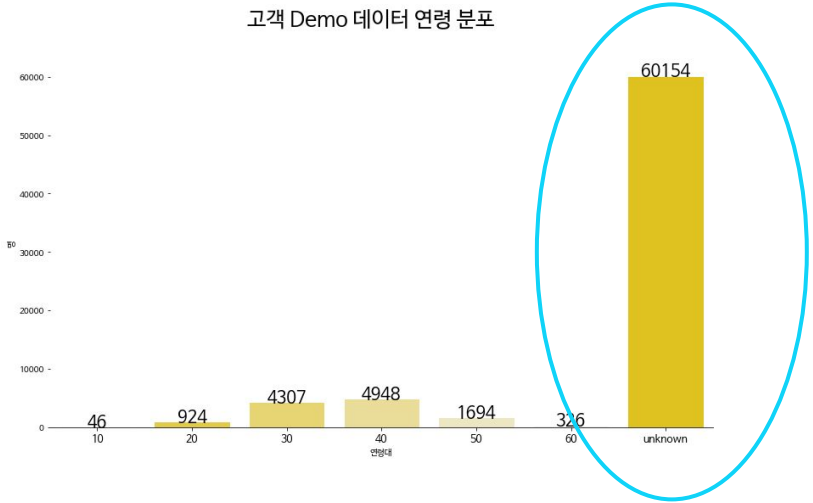
No	변수명(영문)	변수명(국문)	상세설명	PK
1	CLNT_ID	클라이언트ID	고객을 고유하게 식별할 수 있도록 랜덤으로 부여된 ID	Y
2	CLNT_GENDER	성별	성별정보 [남자: M/ 여자: F/ 정보없음 : unknown]	
3	CLNT_AGE	연령대	연령대 정보 [10대이하/ 20대 / 30대 / 40대 / 50대 / 60대이상 / 정보없음: unknown]	

Row: 72399
→ 고객 72399명의 Demo

고객 Demo 데이터 성별 분포



고객 Demo 데이터 연령 분포



데이터셋

L.POINT 온라인 행동정보 데이터

온라인 행동 정보란, 고객의 온라인 행동에 대한 기록으로써
유입부터 구매까지 모든 행동 과정을 분석할 수 있는 데이터

고객	유입	행동	
			
방문자	PC/모바일 통한 접속	사이트 내 클릭, 페이지 뷰 등	장바구니 담기, 상품구매 등

데이터셋

L.POINT 온라인 행동정보 데이터 설명 테이블

No	변수명(영문)	변수명(국문)	상세설명	PK
1	CLNT_ID	클라이언트ID	고객을 고유하게 식별할 수 있도록 랜덤으로 부여된 고객 ID	Y
2	SESS_ID	세션ID	Web/App에 접속 후 세션이 시작될 때 부여된 순번 ID ★하나의 클라이언트ID에 여러 개의 세션 ID가 발급될 수 있음	Y
3	HIT_SEQ	조회일련번호	조회 순서를 알 수 있도록 부여된 일련번호	Y
4	ACTION_TYPE	행동유형	총 8가지의 행동 유형을 구분한 코드 [0.검색/ 1.제품 목록/ 2.제품 세부정보 보기/ 3.장바구니 제품 추가/ 4.장바구니 제품 삭제/ 5.결제 시도 / 6.구매 완료/ 7.구매 환불/ 8.결제 옵션]	
5	BIZ_UNIT	업종단위	온라인 및 오프라인 이용처를 구분하는 단위코드	
6	SESS_DT	세션일자	세션일자 (YYYYMMDD)	
7	HIT_TM	조회시각	조회시각 (HH:MM)	
8	HIT_PSS_TM	조회경과시간	세션이 시작된 이후 해당 조회까지 경과한 시간 (단위: 밀리초)	
9	TRANS_ID	거래ID	구매 내역을 고유하게 식별할 수 있도록 랜덤으로 부여된 ID	
10	SRCH_KWD	거래ID	고객이 검색한 키워드	
11	TOT_PAG_VIEW_CT	검색 키워드	세션 내의 총 페이지(화면)뷰 수	
12	TOT_SESS_HR_V	총페이지조회건수	세션 내 총 시간(단위: 초)	
13	TRFC_SRC	유입채널	고객이 유입된 채널 [DIRECT/PUSH/WEBSITE/PORRTAL_1/PORTAL_2/PORTAL_3/unknown]	
14	DVC_CTG_NM	기기유형	기기 유형 [mobile_web / mobile_app / PC]	

Row:3196362
→ 고객의 세션 별 온라인 행동 기록 데이터가 총 3196362 개

데이터 가공

Input 가공

Sliding Window

CLNT_ID(고객), SESS_ID(세션) 별 HIT_SEQ(조회일련번호)에 따라 정렬한 뒤, 앞의 10 steps를 x, 11번째 step을 y로 정한다. window를 1 step씩 뒤로 이동시키며 이 과정을 반복하여 데이터 생성

예시

CLNT_ID: 2
SESS_ID: 1

action_types

0	1	2	0	0	3	5	5	0	1	2	0	0	0	5

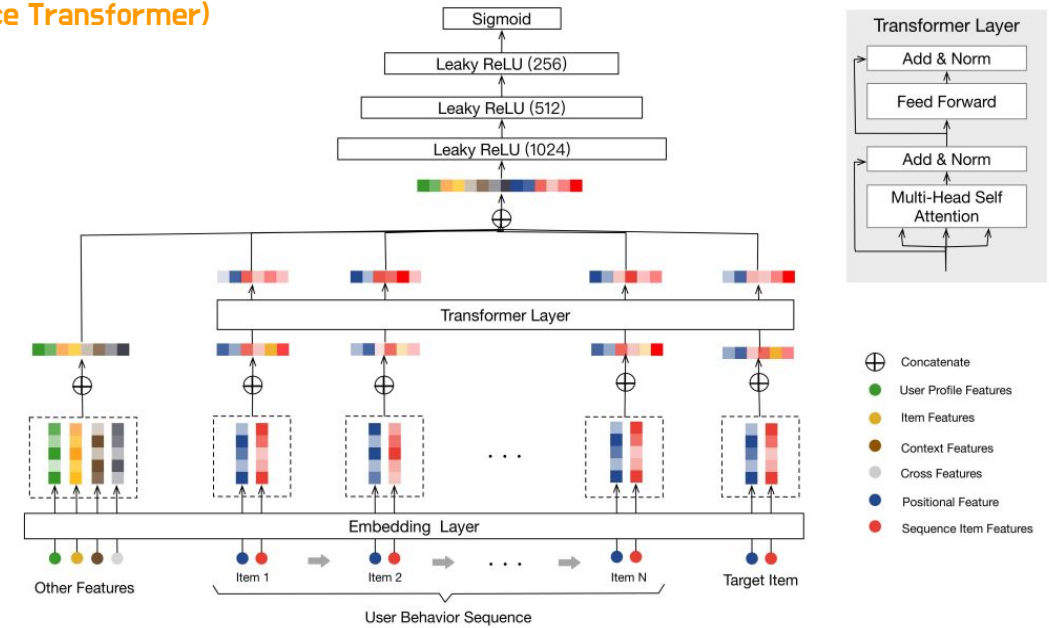
10 steps : x | y

10 steps : x | y

10 steps : x | y

모델 구조

BST (Behavior Sequence Transformer)



참고 논문

Behavior Sequence Transformer for E-commerce Recommendation in Alibaba
- Qiwei Chen, Huan Zhao* Wei Li, Pipei Huang, Wenwu Ou Alibaba Group Beijing & Hangzhou, China

모델 구조

Epochs: 100
Batch_size: 128
Learning rate: 0.0005
Optimizer: Adam
Loss: Categorical Cross Entropy

BST (Behavior Sequence Transformer)

Output(predicted y)
Softmax
FC Layer
모든 Feature Concatenate
Transformer Layer
Feature들끼리 Concatenate
Embedding Layer
Input(x)

FC Layer
- Leaky ReLU(256)
- Leaky ReLU(512)
- Leaky ReLU(1024)

Input
- User Behavior: ACTION_TYPE (x)
- Other Features: HIT_SEQ, BIZ_UNIT, TRFC_SRC, DVC_CTG_NM, WEEKDAY(요일), HOUR(시간)

참고 논문

Behavior Sequence Transformer for E-commerce Recommendation in Alibaba
- Qiwei Chen, Huan Zhao* Wei Li, Pipei Huang, Wenwu Ou Alibaba Group Beijing & Hangzhou, China

3

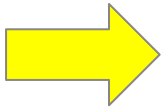
Categorical Data의 시계열 활용

Time Series Application of Categorical Data

Categorical data in time series

1) One hot encoding

DVC_CTG_NM
mobile app
mobile web
pc



mobile app	mobile web	pc
1	0	0
0	1	0
0	0	1

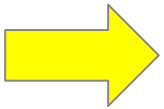
특징

- 카테고리 개수만큼 차원을 갖는 벡터 생성
- 카테고리 개수가 커지면 데이터가 굉장히 sparse 해짐
- 카테고리 벡터 간에 거리가 동일하므로 관계를 분석하기 어려움

Categorical data in time series

2) Embedding

DVC_CTG_NM	index
mobile app	1
mobile web	2
pc	3



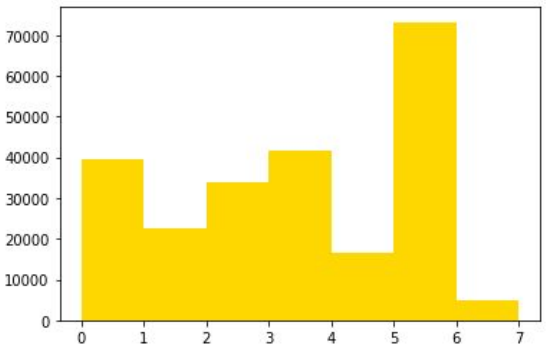
index	Look up table (vector)
1	D차원
2	
3	

특징

- 각 카테고리에 대해 원하는 차원 수(D)만큼 설정할 수 있음 (sparse 방지)
- Embedding 값이 학습을 통해 의미적으로 비슷한 변수들은 군집이 되는 효과가 있음

예측 결과

	Test set
Accuracy (%)	61.93
F1 Score	61.00

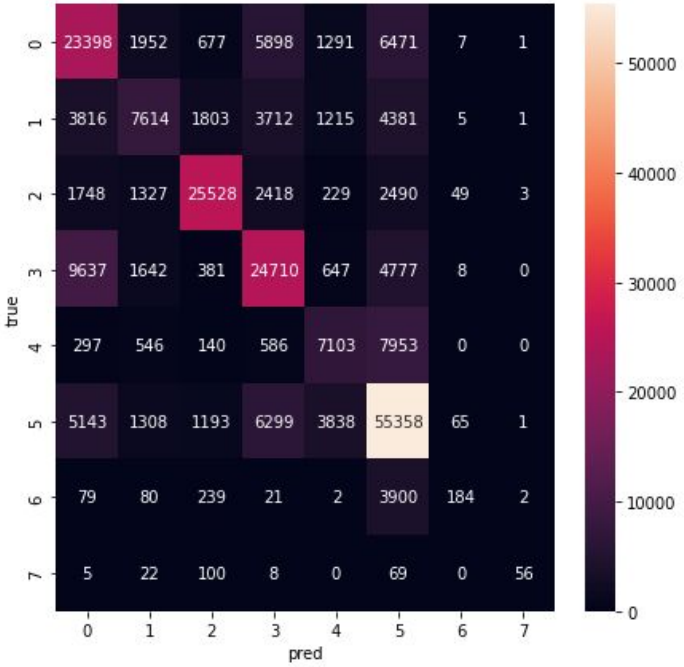


Class	행동	비율(%)
0	검색	17.08
1	제품 목록	9.7
2	제품 세부정보	14.54
3	장바구니 추가	17.98
4	장바구니 삭제	7.15
5	결제 시도	31.5
6	구매 완료	1.94
7	구매 환불	0.11

이전 10개의 sequence를 활용.
다음 1개의 행동 예측 결과.

Test set 기준 약 62%의 정확도로.
레이블의 분포 및
인간의 행동을 예측했다는 점을 고려하면
괜찮은 예측이라고 볼 수 있음.

Confusion matrix



Confusion Matrix를 살펴보면
Class 간 불균형으로 인한 문제점이 드러남.

즉 샘플이 많은 Class일 수록
더 높은 정확도를.
샘플이 적은 Class일 수록
더 낮은 정확도를 보임.

Plot: Embedded categories

학습된 모델에서

Embedding layer만 가져온 후,

데이터를 넣으면

임베딩된 벡터를 가져올 수 있음.

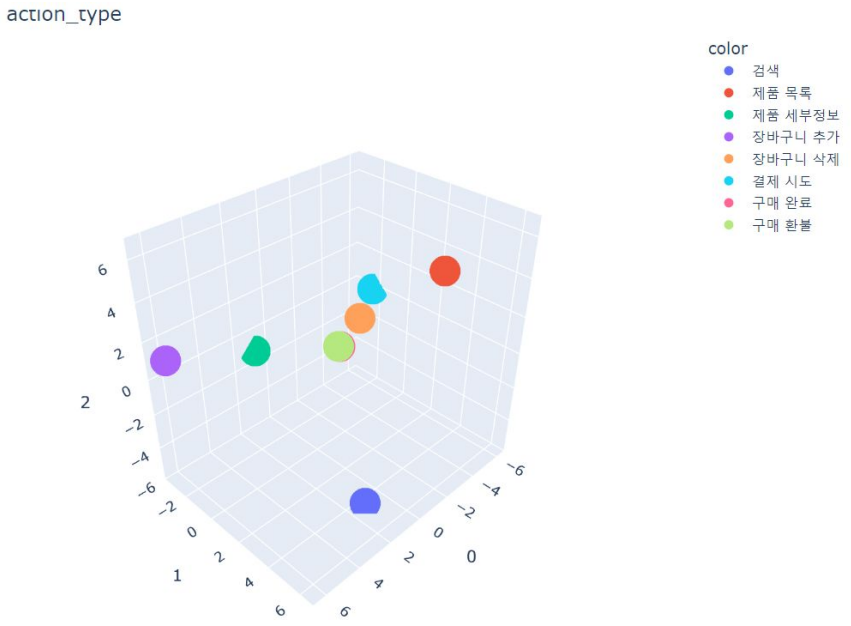
action_type을 포함.

다양한 범주형 데이터를 임베딩한

64차원의 벡터를

3차원으로 차원 축소한 후

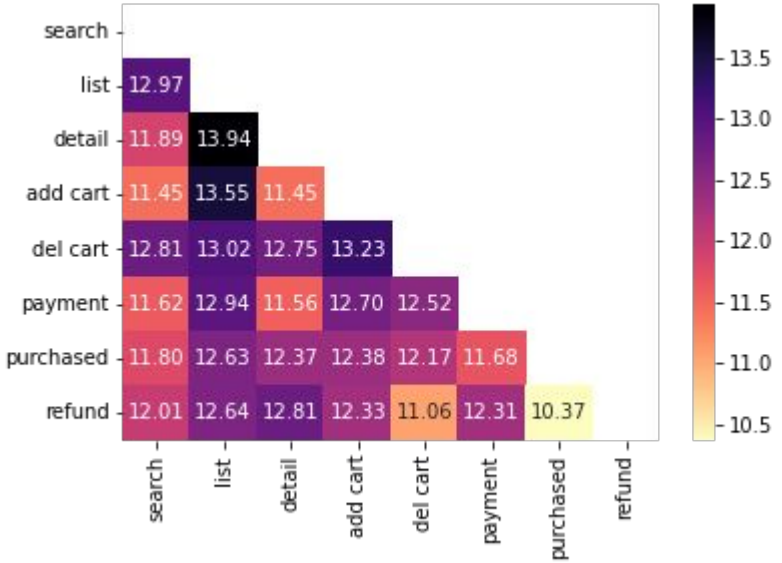
각 벡터의 위치를 표현



Euclidean distance matrix

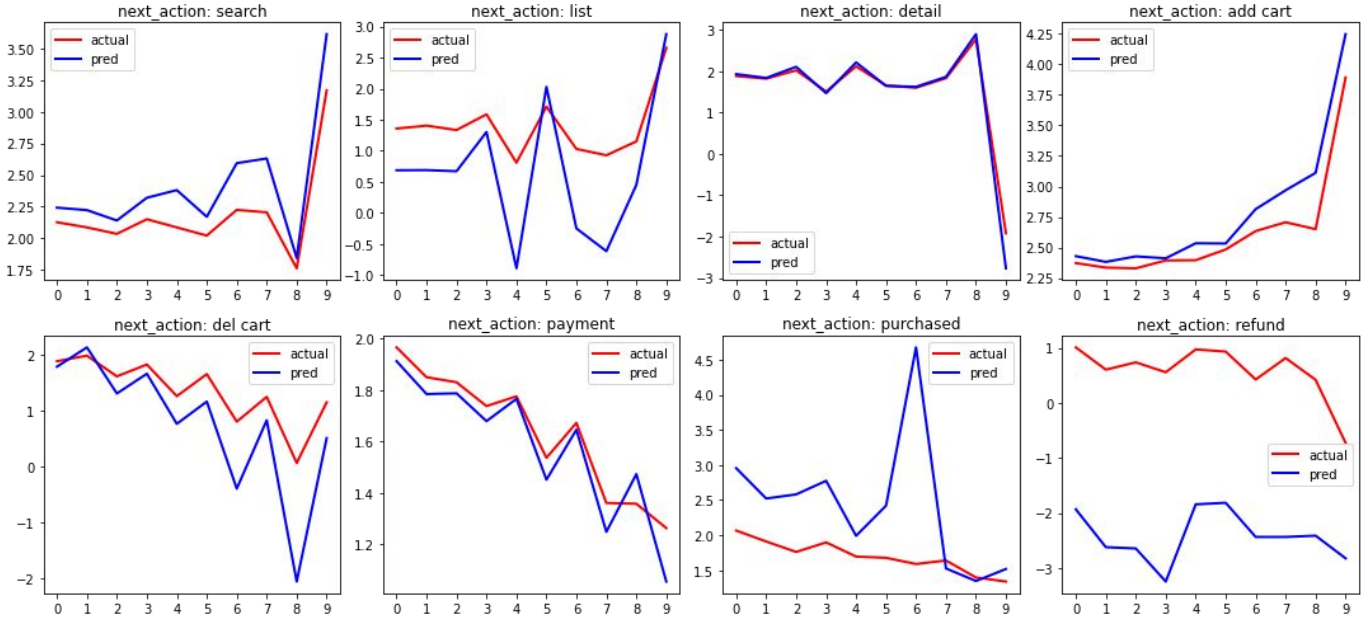
각 action_type 끼리의
유클리디안 거리를 구한 후
관계를 파악.

거리가 가까울 수록
서로 연관된 행동이라고
판단할 수 있음.



클래스 별 입력 시퀀스 비교

action_type	1
list	-5.198564
del cart	-3.663511
refund	-3.335166
purchased	-1.823763
payment	1.461432
search	2.659162
add cart	4.622114
detail	5.278297



4 모델 튜닝

Model Tuning

Model Tuning 필요성

낮은 정확도를 향상시킬 수 있는 방안 필요

Action Type을 **군집화**시키면 어떨까?

Model Tuning 유형

Tuning 0 기존의 8가지 Action Type 유지

[0.검색 / 1.제품 목록 / 2.제품 세부정보 보기 / 3.장바구니 제품 추가 /
4.장바구니 제품 삭제 / 5.결제 시도 / 6.구매 완료 / 7.구매 환불 / 8.결제 옵션]

Tuning 1 직관적인 분류

0, 1, 2, 5 -> 제품 관심도(1)

3, 6 -> 수익 관련 변수(2)

4, 7 -> 비용 변수(3)

Tuning 2 임베딩 기반 분류

-> 군집 1

-> 군집 2

-> 군집 3

Model Tuning 유형

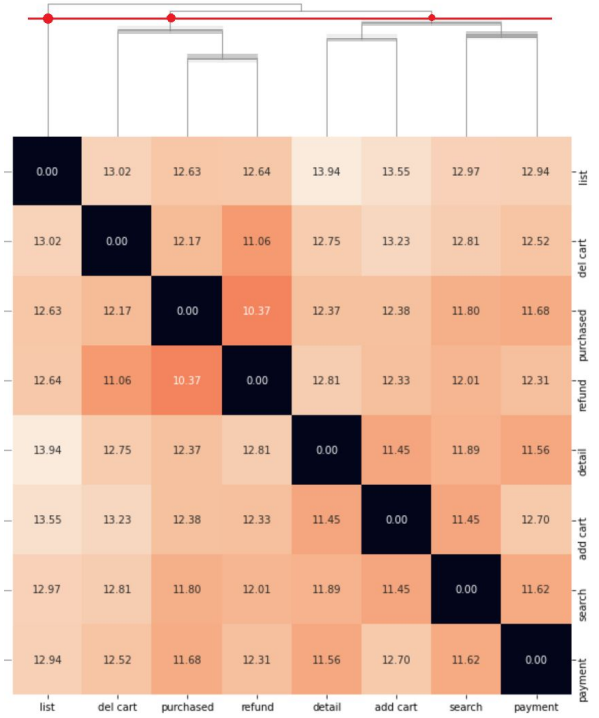
Tuning 2 임베딩 기반 분류

-> 군집 1: 제품 목록

-> 군집 2: 장바구니 제거, 구매 완료, 구매 환불

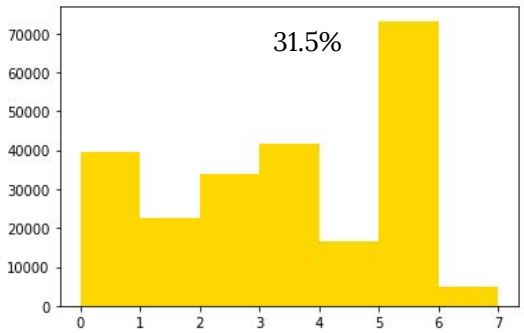
-> 군집 3: 제품 세부정보, 장바구니 추가, 검색, 결제 시도

클래스별 임베딩된 벡터의 유클리디안 거리 행렬과
계층적 클러스터링을 이용하여 그룹화

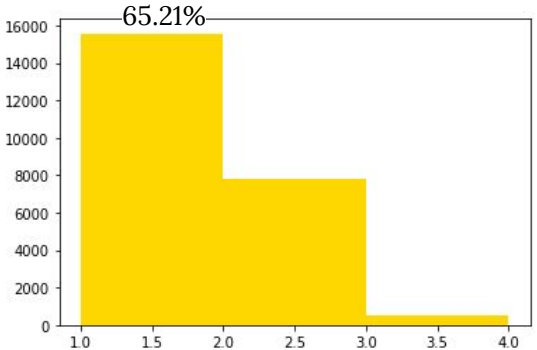


Model Tuning 결과

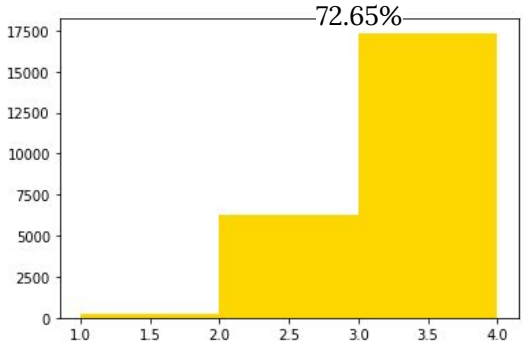
	Tuning 0	Tuning 1	Tuning 2
Accuracy (%)	61.93	73.6	81.64



Tuning 0



Tuning 1



Tuning 2

5

모델 기대효과

Expected Effectiveness

모델 기대효과

知彼知己 百战不殆

상대를 알고 나를 알면, 백번 싸워도 위태롭지 않다

- Cold Start 문제

: 고객의 실시간 행동패턴으로 다음 행동 예측 가능 (Generalized Prediction)

: 신규고객에 대한 선호도 파악 시간 절약

: 쌓인 신규고객의 행동 패턴은 재학습하여 정확도 높임 (Personalized Prediction)

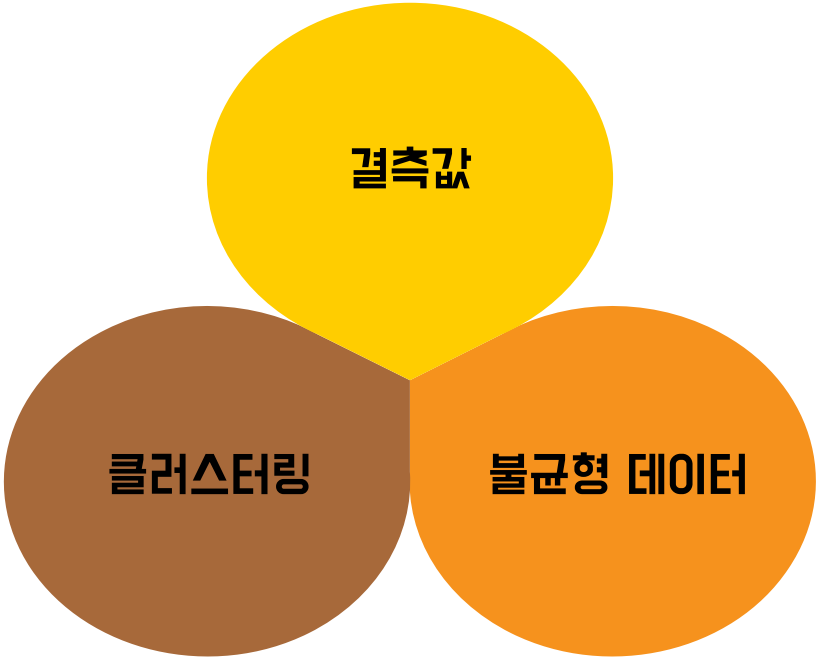
- Marketing 비용 절약

: 광고 노출 로직의 개인화 가능

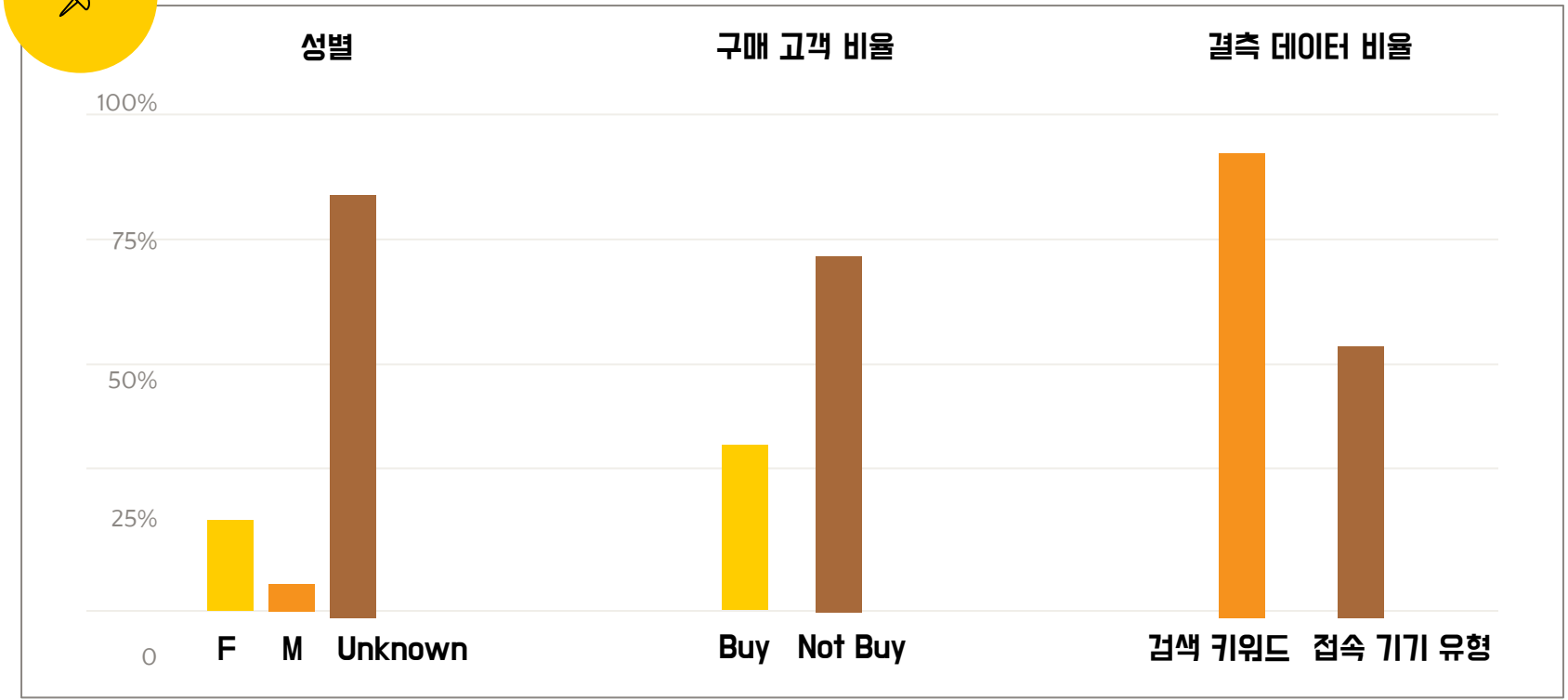
6 한계

Limit

한계



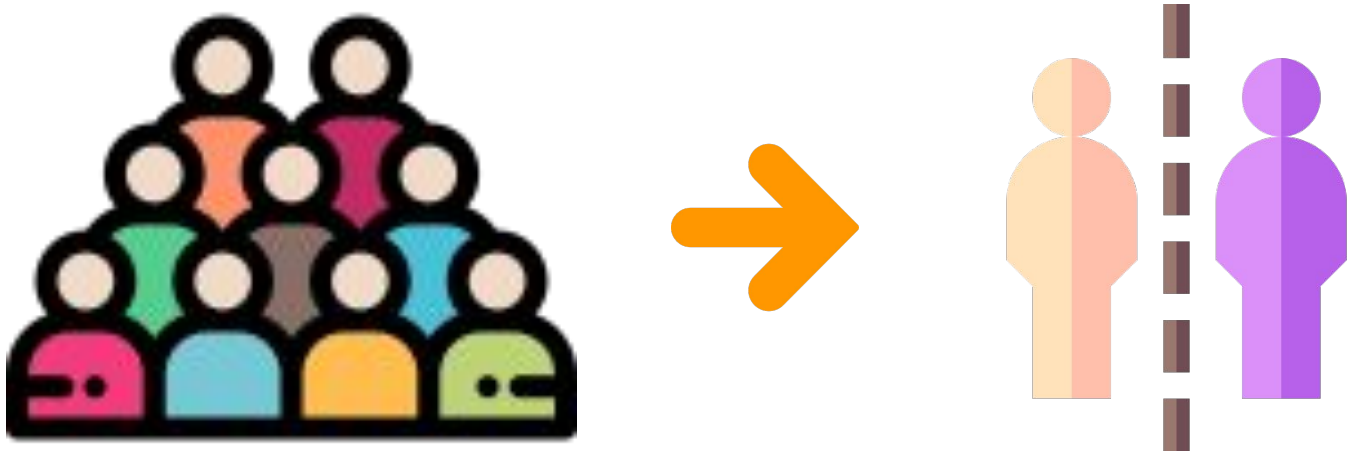
결측값



Demographic 데이터는 절반 이상이 결측값.
거래 데이터는 구매자에 대한 데이터만 있지만 구매자의 비율이 적음.
서치 키워드, 접속 기기유형 등 결측값으로 인해 사용 못 한 변수들 다수 존재.

클러스터링

고객 분류의 필요성

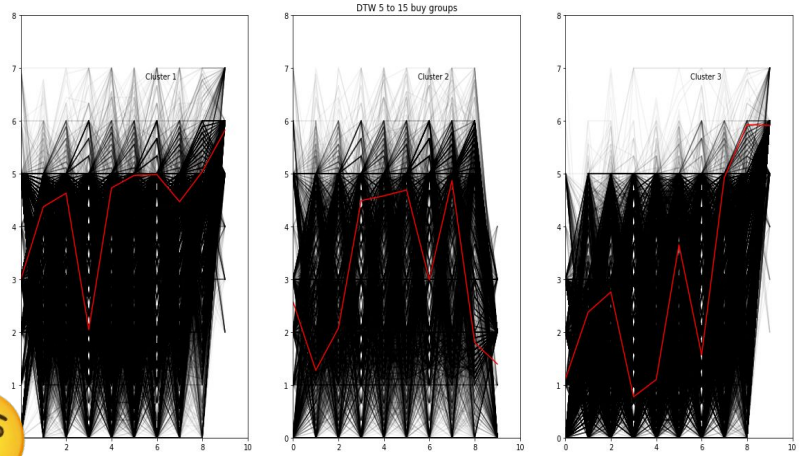


다양한 특성을 가진 전체 고객을 한 결과값으로 일반화하는 것은 무리.
비슷한 고객끼리 분류하는 것이 필요.

클러스터링

복잡한 데이터

군집별 경향성이 있다고 판단 어려움.



범주형 데이터

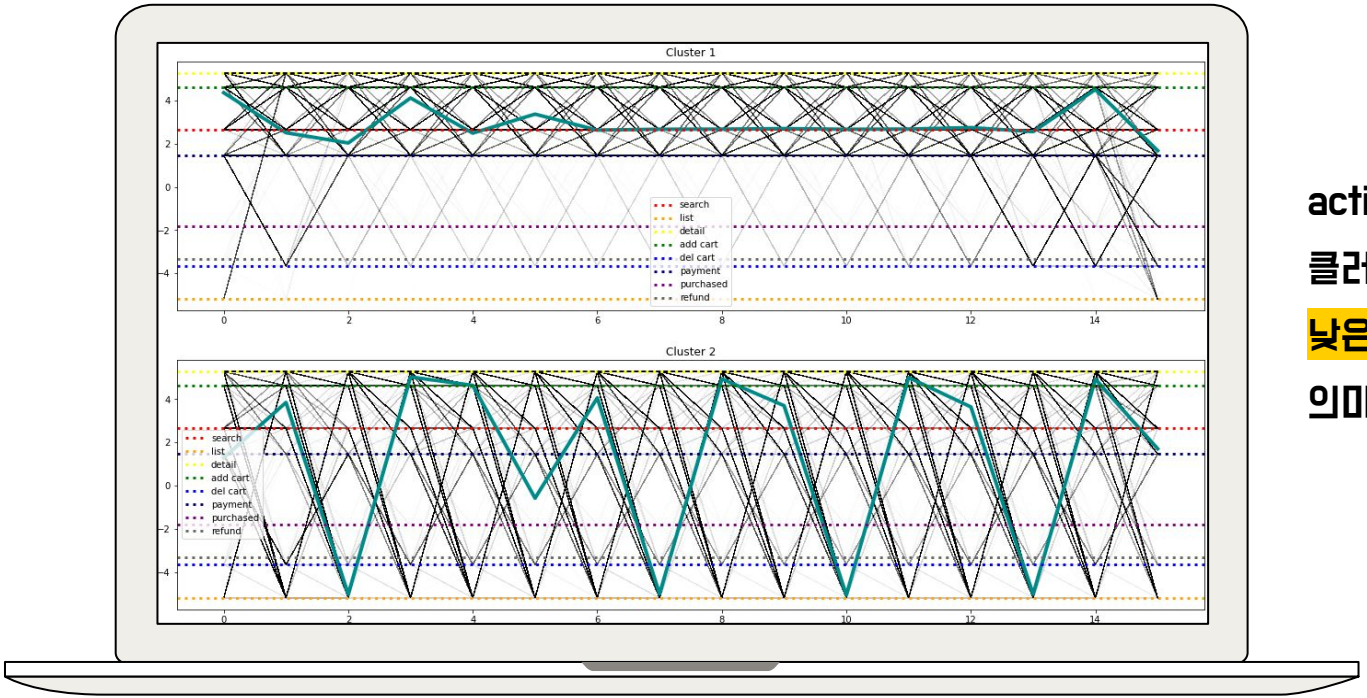
범주형 데이터를 DTW kmeans를 통해 군집화할 수 없음.

1 → 제품목록

4 → 장바구니, 제품삭제

7 → 구매 환불

클러스터링



action_type을 임베딩 한 후

클러스터링을 시도해도.

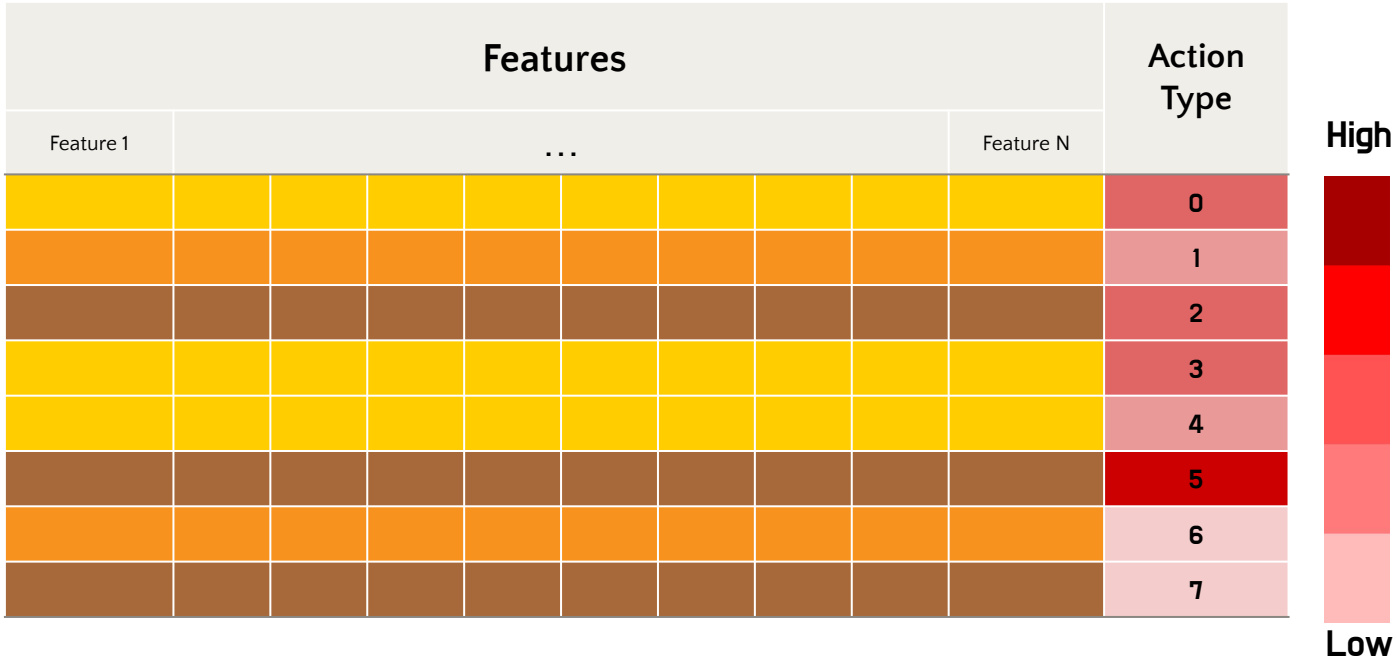
낮은 실루엣 계수 등으로

의미있는 결론을 이끌 수는 없었음.

불균형 데이터

불균형 데이터

Action type 간 비율 차이 많이 남



Let's review some concepts

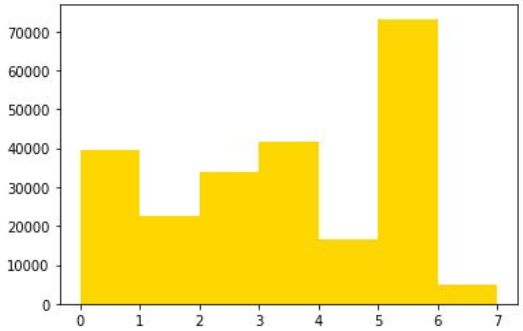
층화추출

모집단을 action type의 수(8개)의 층만큼 나눈 후 각 층에서 일정 표본 크기만큼 표본 추출하는 방법.

However...

Action type의 비율 간 극심한 차이가 남.

데이터의 비율이 현저히 적은 action type의 경우 매우 낮은 정확도를 보임.



Thanks!

Any questions ?

DSL

- 김지오. 박준우. 이승재. 장윤태. 조수연. 조영규