

머신러닝 기초 (파이썬) 추천 알고리즘 적용 사례 소개

목 차

1. 빅데이터 분석이란?

2. 빅데이터 분석 툴 소개

3. Python기초 & Classification

4. Data analysis in Marketing

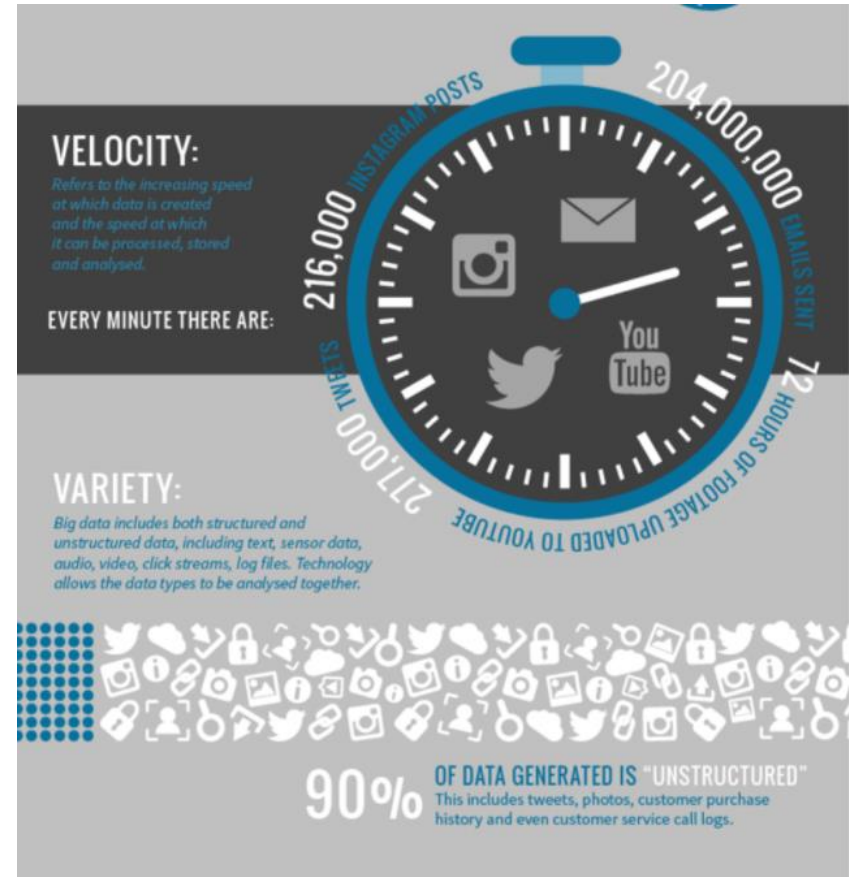
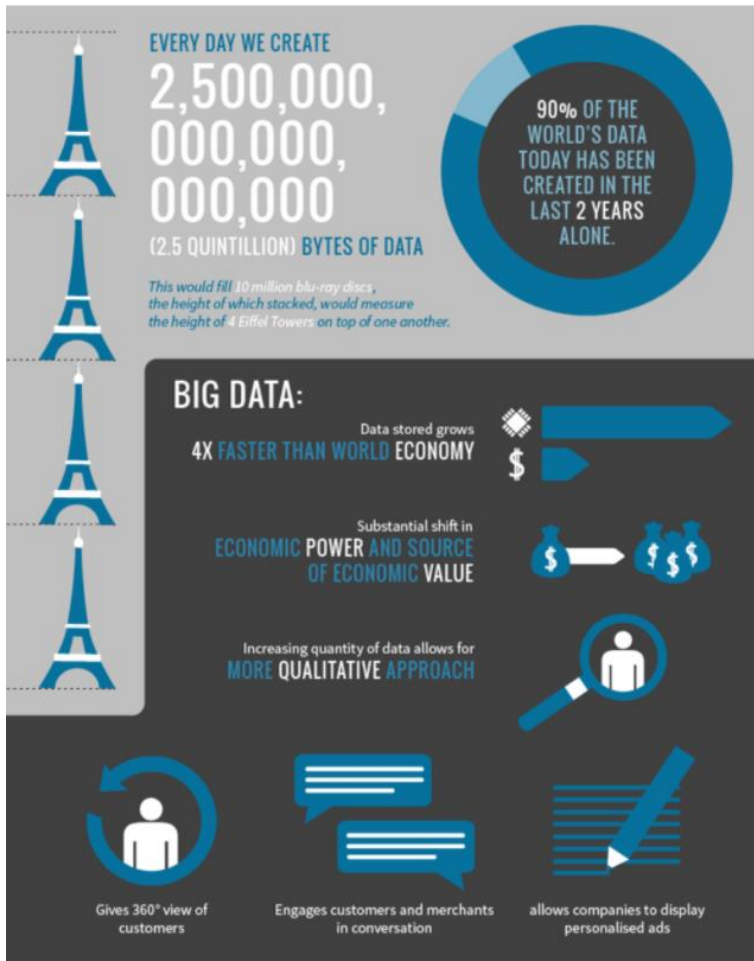
5. 개인화 추천 시스템

Q & A

1. 빅데이터 분석이란?



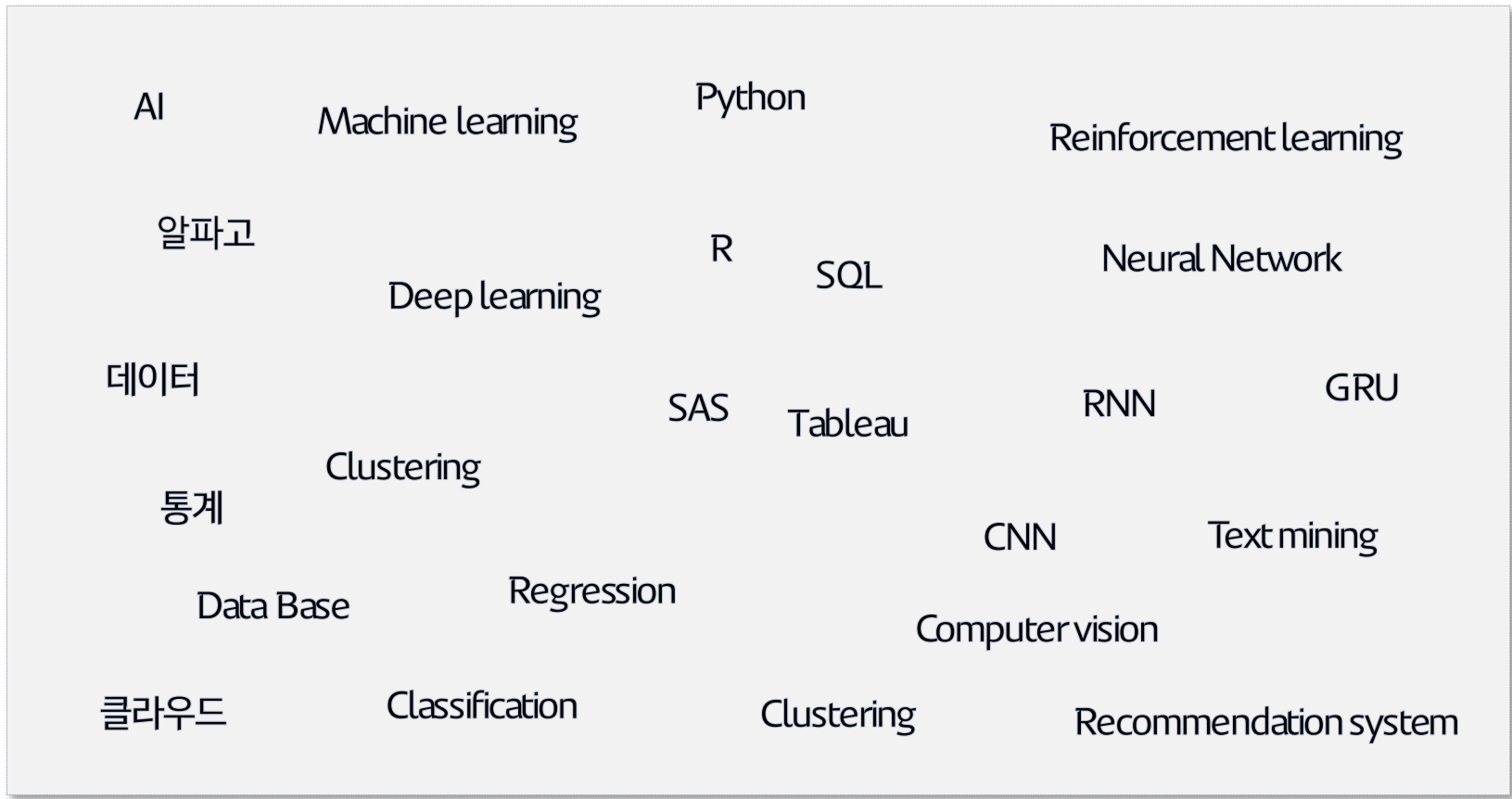
1. 빅데이터 분석이란?



90 Percent of the Big Data We Generate Is an Unstructured Mess, by Eric Griffith, 2018
<https://www.pcmag.com/news/364954/90-percent-of-the-big-data-we-generate-is-an-unstructured-mess>
<https://www.vouchercloud.com/resources/everyday-big-data>

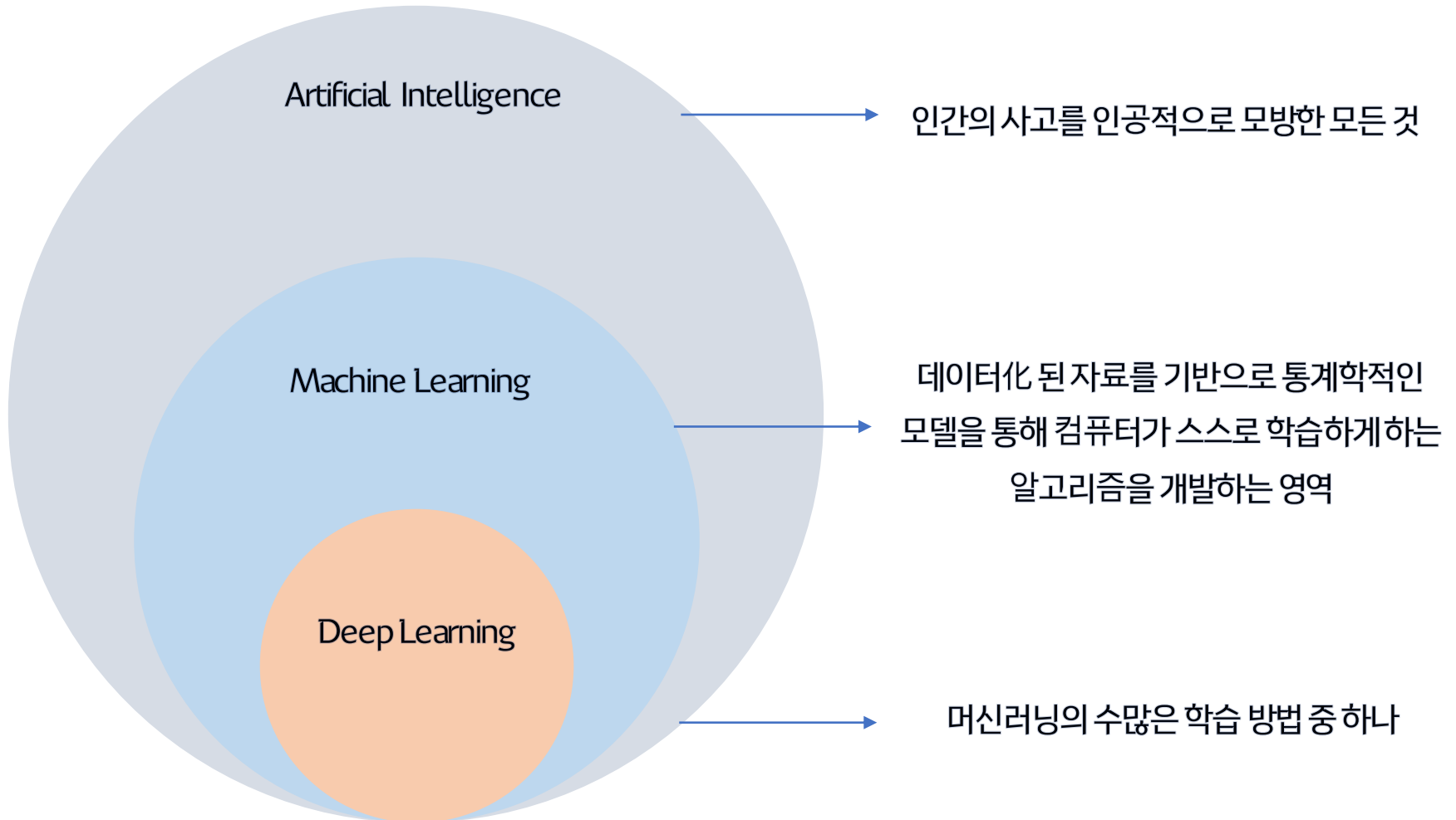
1. 빅데이터 분석이란?

빅데이터 분석



1. 빅데이터 분석이란?

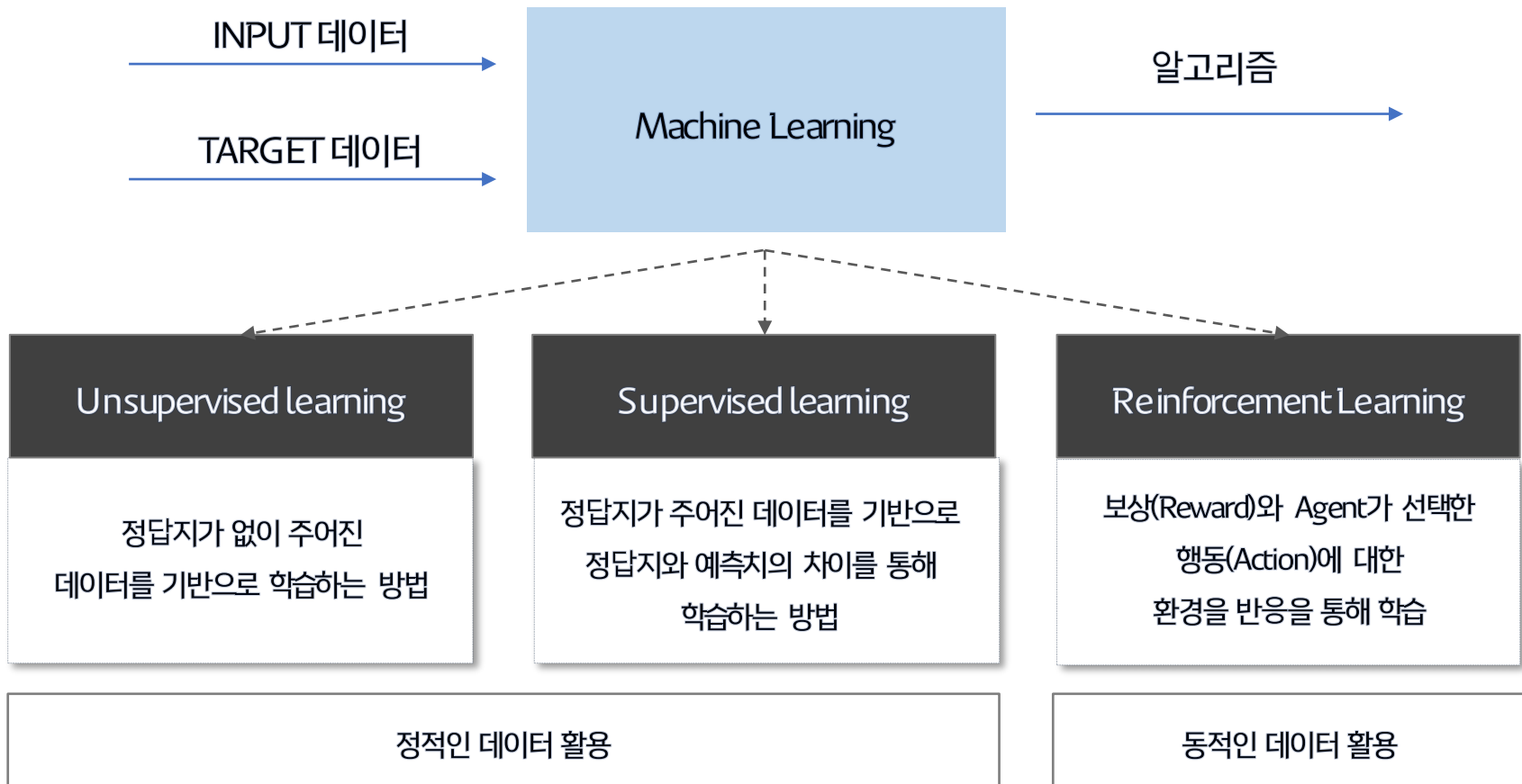
빅데이터를 활용 방법



1. 빅데이터 분석이란?

Machine Learning

AI의 한 범주로서 컴퓨터가 스스로 학습하게 하는 알고리즘을 개발하는 분야

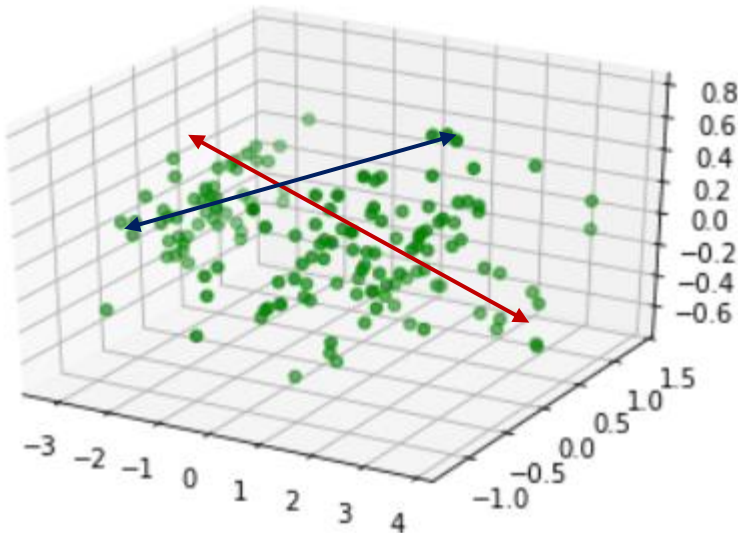


1. 빅데이터 분석이란?

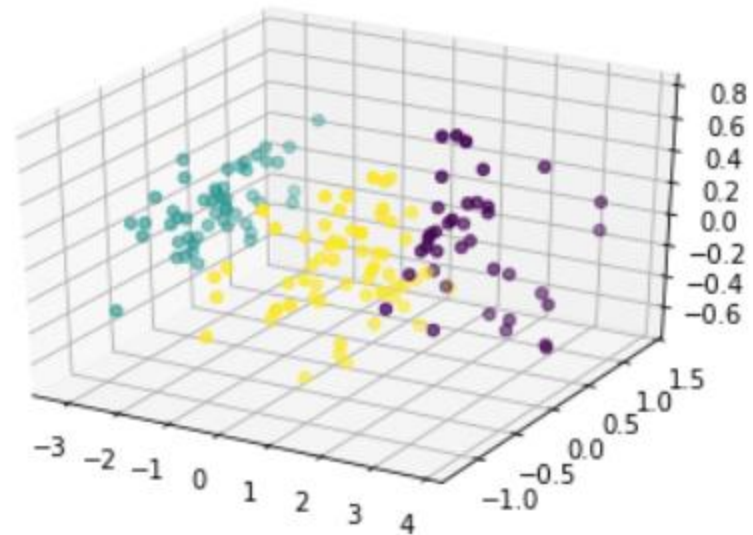
Machine Learning

Unsupervised learning

정답지가 없이 주어진 데이터를 기반으로 학습하는 방법



미 분류된 데이터 속에 숨겨진 구조를 찾아내는 과정



Kmeans++, Kprototype, GMM, DBSCAN
PCA, LDA ...

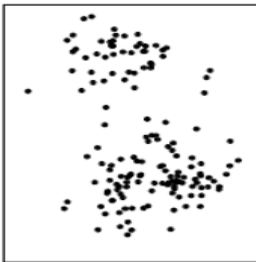
1. 빅데이터 분석이란?

Machine Learning

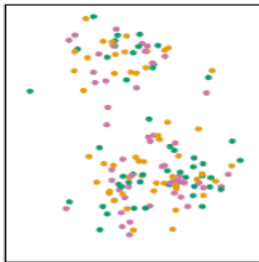
Unsupervised learning

정답지가 없이 주어진 데이터를 기반으로 학습하는 방법

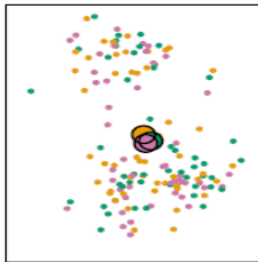
[k-Means 알고리즘 예시]



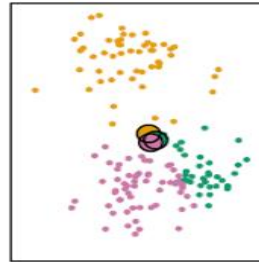
- Step 1.
- 데이터 좌표화



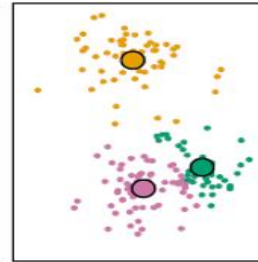
- Step 2.
- 클러스터 개수(k)설정
 - 데이터에 클러스터 무작위 할당



- Step 3.
- 설정한 각 클러스터 개수(k)에 대한 mean 값 산출



- Step 4.
- 각 mean값과, 모든 데이터 포인트들과의 거리 계산 후 클러스터 재할당
 - 재할당된 각 클러스터의 mean값 산출



- Step 5.
- 더 이상의 클러스터 재할당이 없을 때까지 Step 4의 과정을 반복

Kmeans++, Kprototype, GMM, DBSCAN

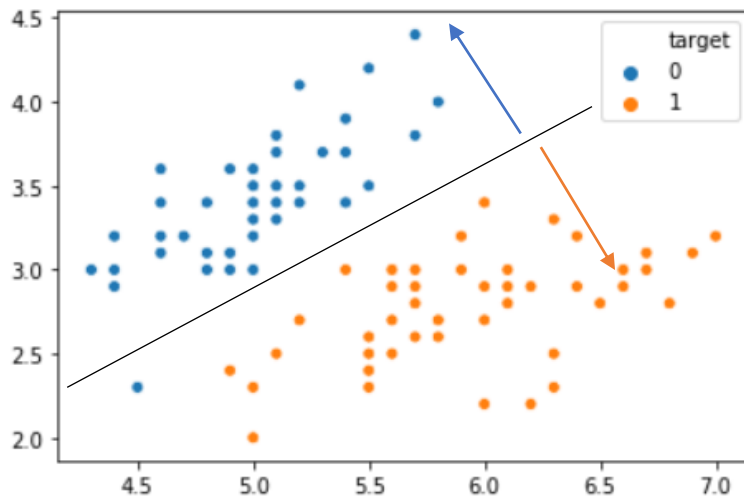
PCA, LDA ...

1. 빅데이터 분석이란?

Machine Learning

supervised learning

정답지가 주어진 데이터를 기반으로 정답지와 예측치의 차이를 통해 학습하는 방법



Classification (분류)

주어진 데이터의 Feature & Label값을 학습

> 생성된 모델을 통해 새로운 데이터 값이 주어질 때 값을 예측하는 분석 방법

Regression (회귀)

주어진 데이터의 Feature & 결정값을 학습

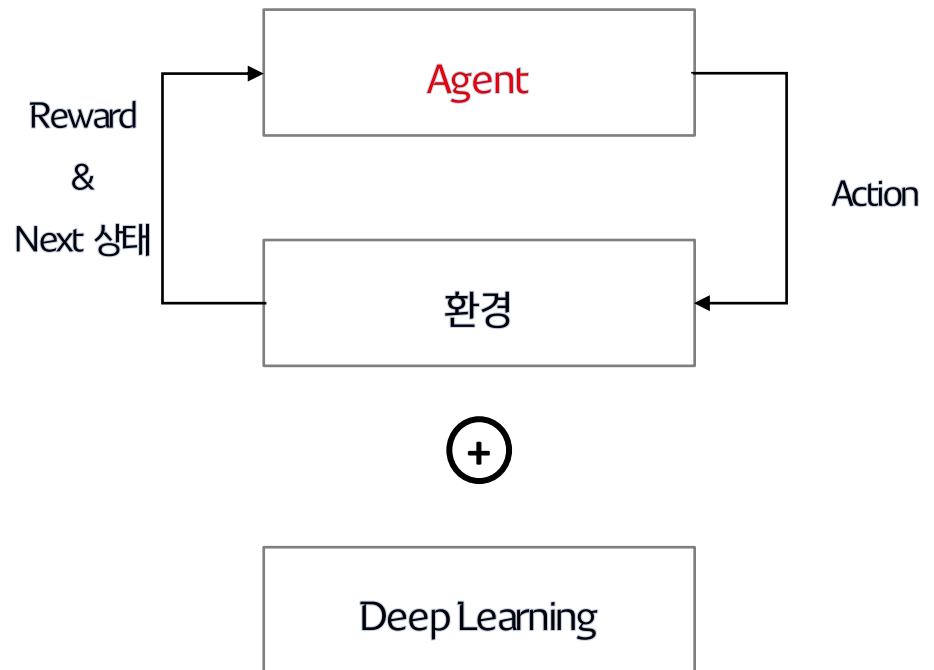
“경사하강법”을 통해 최적의 회귀 계수를 찾는 방법

1. 빅데이터 분석이란?

Machine Learning

Reinforcement learning

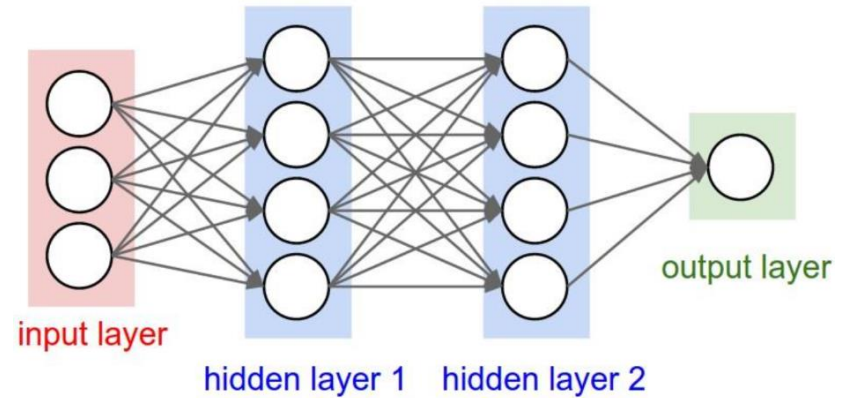
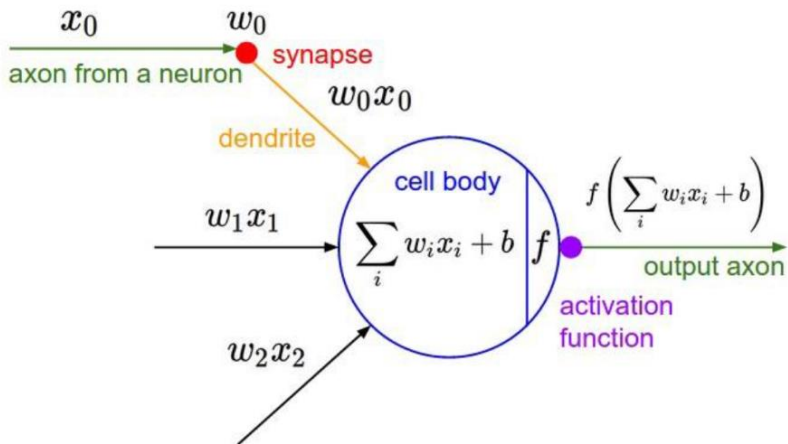
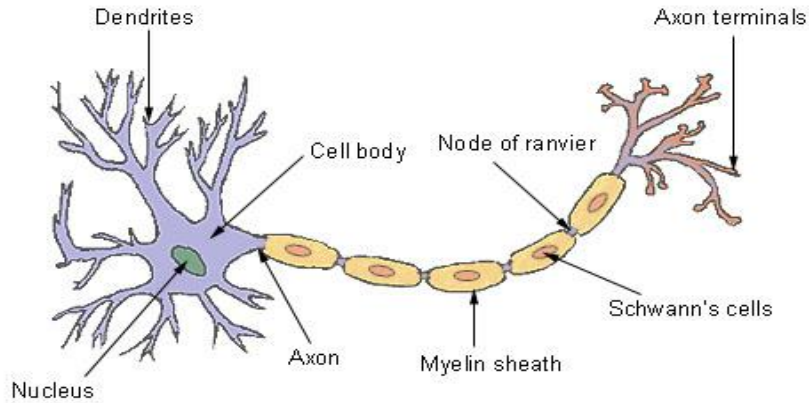
보상(Reward)와 Agent가 선택한 행동(Action)에 대한 환경을 반응을 통해 학습



1. 빅데이터 분석이란?

답러닝이란?

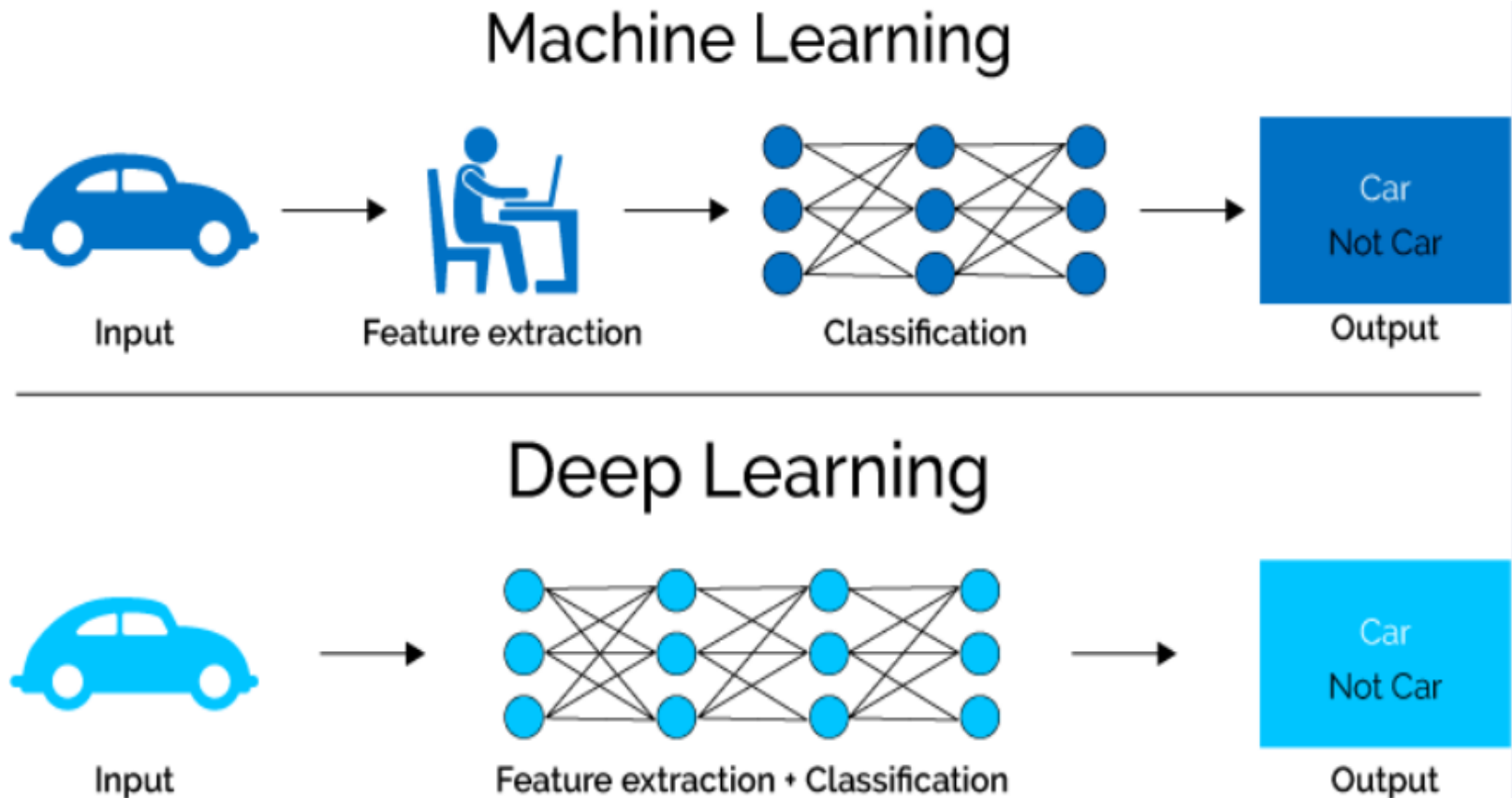
Structure of a Typical Neuron



현재 내가 틀린정도를 '미분(기울기)'한 후 곱하고, 더하고를
역방향으로 반복하며 업데이트

1. 빅데이터 분석이란?

딥러닝이란?



Source: [XenonStack](https://www.xenonstack.com)

2. 빅데이터 분석 툴 소개



2. 빅데이터 분석 툴 소개

빅데이터 분석 툴

R

Python

SAS

But why is Python the most popular?

- Open Source
- High Utility in Statistics and Data Analysis
- Text Analysis
- A Powerful Community
- Concise Legibility
- Various Libraries

Jupyter Notebook

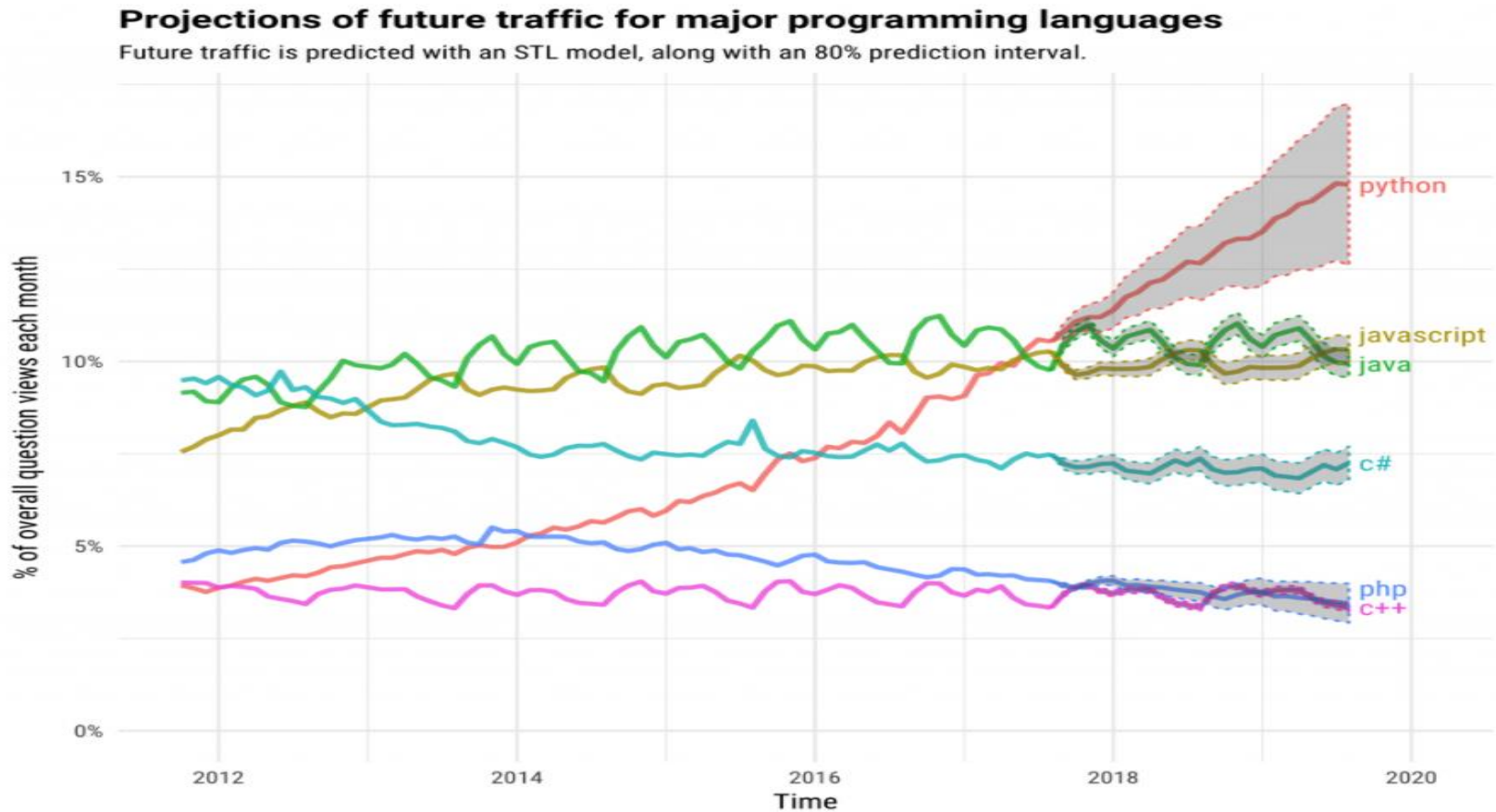
Pycharm

Colab

Sagemaker

2. 빅데이터 분석 툴 소개

빅데이터 분석 툴

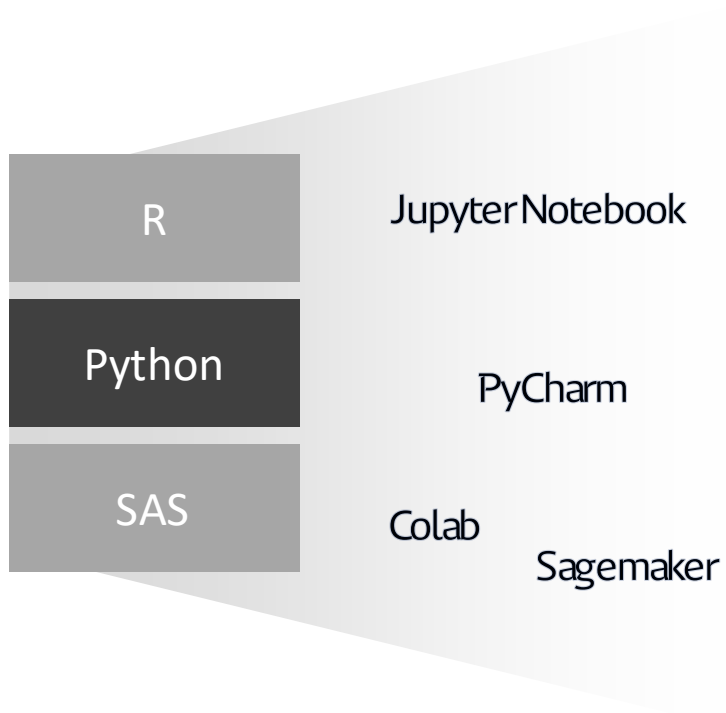


The Incredible Growth of Python, by David Robinson, 2017

<https://stackoverflow.blog/2017/09/06/incredible-growth-python/>

2. 빅데이터 분석 툴 소개

| 연구 방향



Array형태 / 통계수치	Numpy, Scipy
데이터 핸들링	Pandas
데이터 시각화	Matplotlib, Seaborn Plotly
Machine Learning	Sklearn
Deep Learning	Pytorch, Tensorflow Keras,

2. 빅데이터 분석 툴 소개

Jupyter notebook

<https://www.anaconda.com/distribution/>

- Click your OS type – Windows/MacOS/Linux
- Download installer

Anaconda Installers

Windows 

Python 3.8

64-Bit Graphical Installer (466 MB)


32-Bit Graphical Installer (397 MB)

MacOS 

Python 3.8

64-Bit Graphical Installer (462 MB)

64-Bit Command Line Installer (454 MB)

Linux 


Python 3.8

64-Bit (x86) Installer (550 MB)

64-Bit (Power8 and Power9) Installer (290 MB)

2. 빅데이터 분석 툴 소개

Jupyter notebook

 jupyter

QuitLogout

FilesRunningClusters

Select items to perform actions on them.

UploadNew ↕ ↺

☐ 0 ▾

📁 /

Name ▾

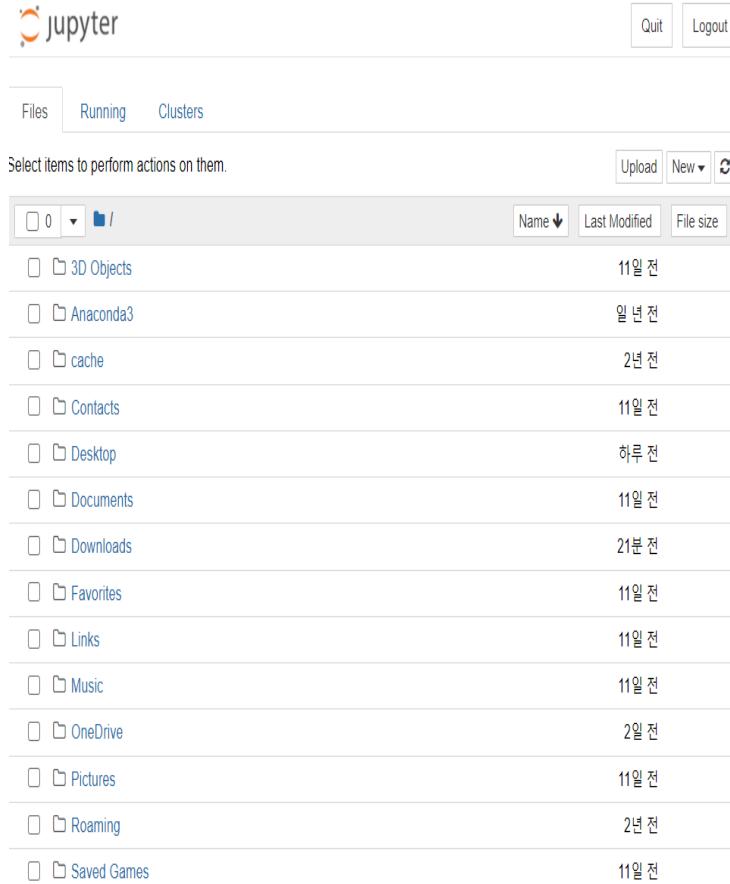
Last Modified

File size

<input type="checkbox"/>	📁 3D Objects	11일 전
<input type="checkbox"/>	📁 Anaconda3	일 년 전
<input type="checkbox"/>	📁 cache	2년 전
<input type="checkbox"/>	📁 Contacts	11일 전
<input type="checkbox"/>	📁 Desktop	하루 전
<input type="checkbox"/>	📁 Documents	11일 전
<input type="checkbox"/>	📁 Downloads	21분 전
<input type="checkbox"/>	📁 Favorites	11일 전
<input type="checkbox"/>	📁 Links	11일 전
<input type="checkbox"/>	📁 Music	11일 전
<input type="checkbox"/>	📁 OneDrive	2일 전
<input type="checkbox"/>	📁 Pictures	11일 전
<input type="checkbox"/>	📁 Roaming	2년 전
<input type="checkbox"/>	📁 Saved Games	11일 전

2. 빅데이터 분석 툴 소개

Jupyter notebook



Python에서 대표적인 대화형으로 분석 가능한 툴

```
In [4]: # Load Dataset
import pandas as pd
train_dataset = pd.read_csv('mnist_train.csv')
test_dataset = pd.read_csv('mnist_test.csv')
```

```
In [5]: train_dataset.head()
```

Out[5]:

	label	1x1	1x2	1x3	1x4	1x5	1x6	1x7	1x8	1x9	...	28x19	28x20	28x21	28x22	28x
0	5	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	4	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
3	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
4	9	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

5 rows × 785 columns

2. 빅데이터 분석 툴 소개

Google – colab



drama_analysis_3.ipynb ☆

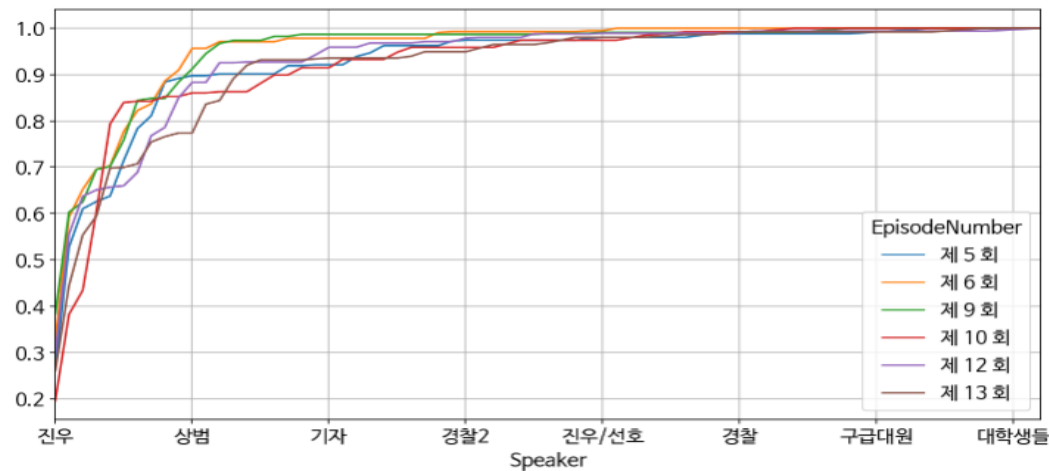
파일 수정 보기 삽입 런타임 도구 도움말 변경사항이 저장되지 않음

+ 코드 + 텍스트 드라이브로 복사

```
[ ] episode_sum_ds=act_episode_ds.sum()  
act_episode_ds = (act_episode_ds/episode_sum_ds)  
act_episode_ds['all'] = act_episode_ds.sum(axis=1)
```

```
## 주요 인물 필터링  
act_episode_ds.sort_values(by=['all'], ascending=False)##  
.cumsum()[['제 5 회', '제 6 회', '제 9 회', '제 10 회', '제 12 회', '제 13 회']].plot(figsize=(12,6))  
plt.grid(True)  
  
actor_list = set()  
for (idx, ep) in act_episode_ds[['제 5 회', '제 6 회', '제 9 회', '제 10 회', '제 12 회', '제 13 회']].iteritems():  
    ep = ep.sort_values(ascending=False, inplace=False).cumsum()  
    actor_list.update(ep[ep<0.7].index.values)  
print(actor_list)
```

{'양주', '민주', '정훈', '진우', '병준', '희주'}



3. Python기초 & Classification

3. Python기초 & Classification

| 모델 학습을 위한 데이터 셋 분리



머신러닝 알고리즘 학습을 위해 사용
데이터의 Feature와 예측값 모두를 가지고 있음

데이터 Feature만을 제공
학습 데이터와는 별도 데이터여야 함

Validation : 학습한 모델의 정확도를 평가하며 Overfitting 을 막기 위함

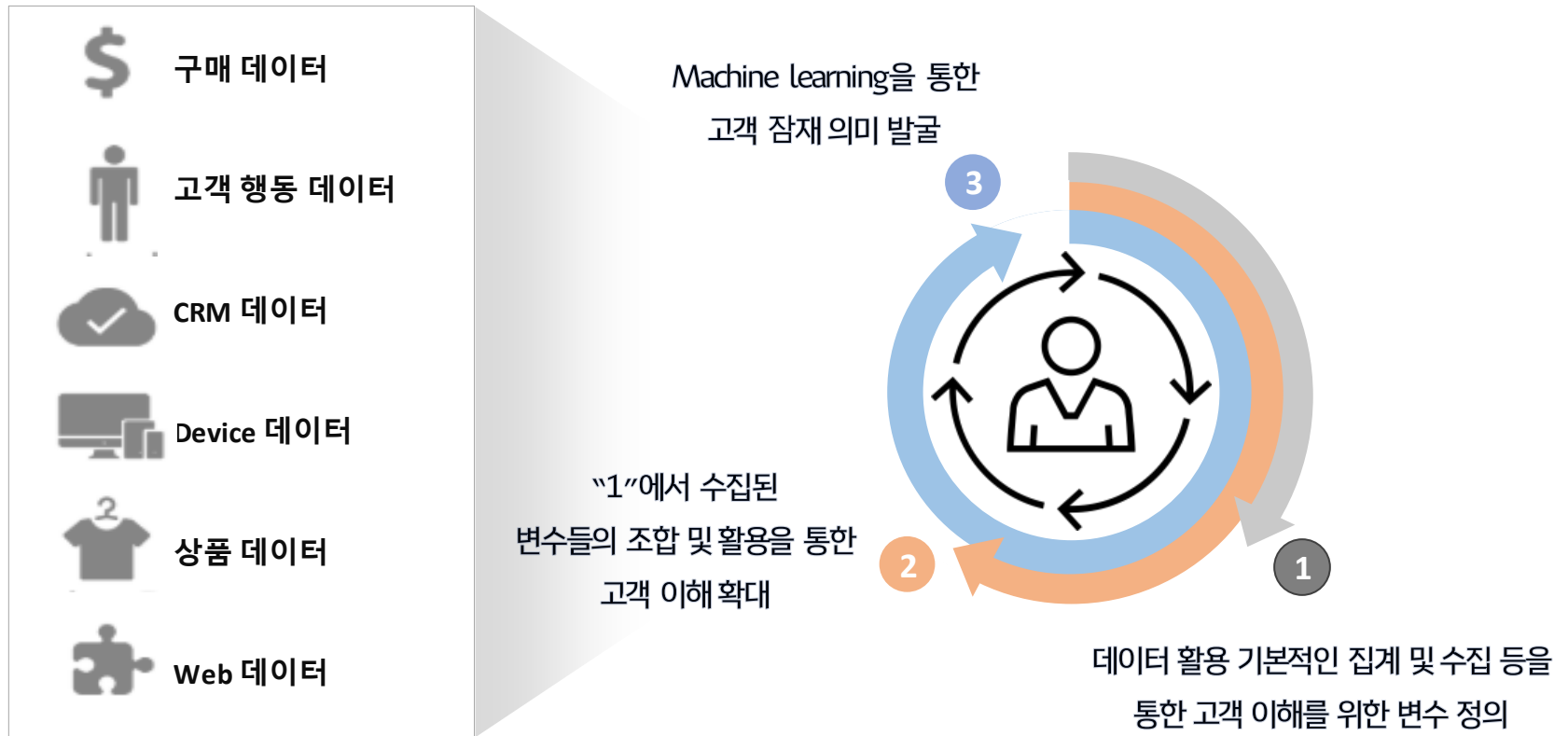
4. Data analysis in Marketing



4. Data analysis in Marketing

고객 이해를 위한 데이터 모델 설계

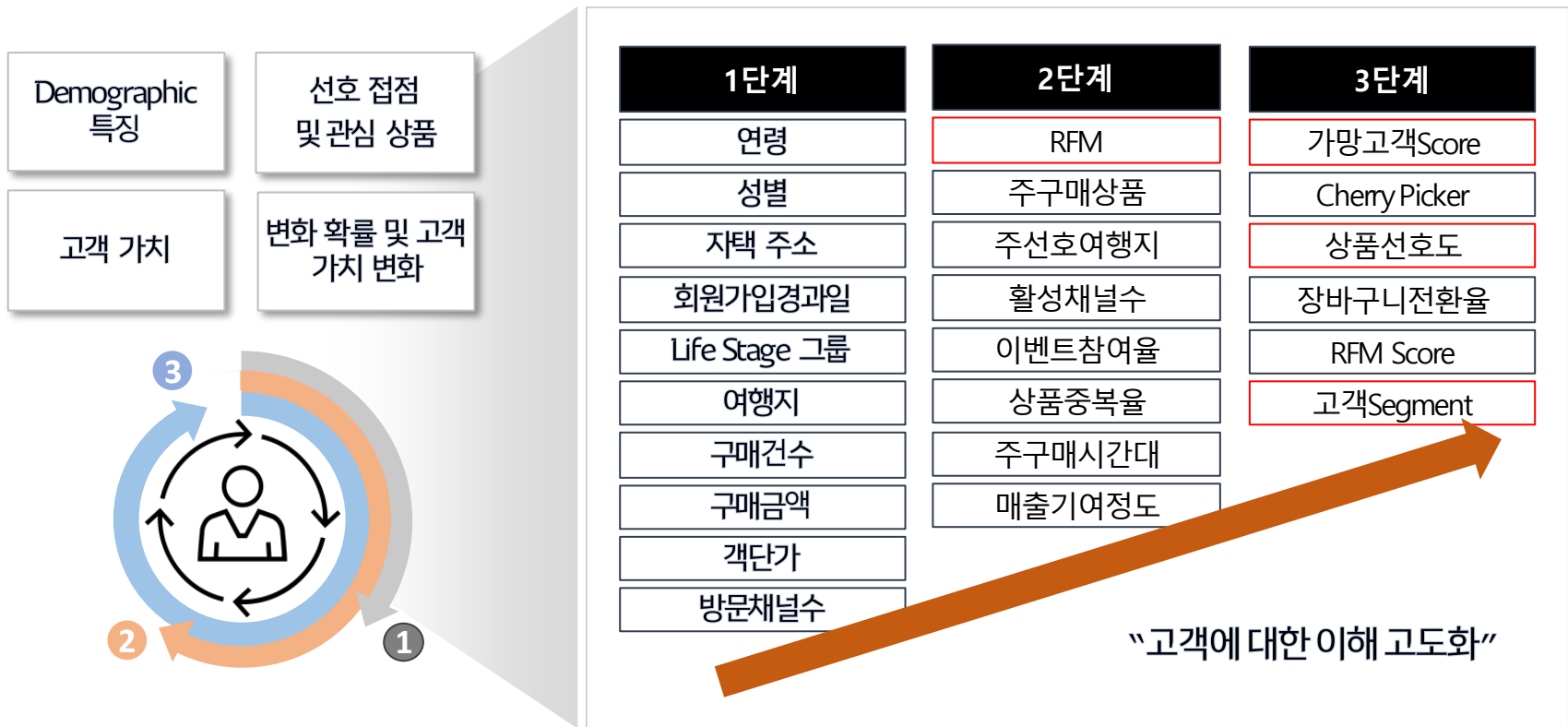
고객 Life style 및 행동 style을 정의할 수 있는 변수 개발 “고객 다각적 이해” 추구



4. Data analysis in Marketing

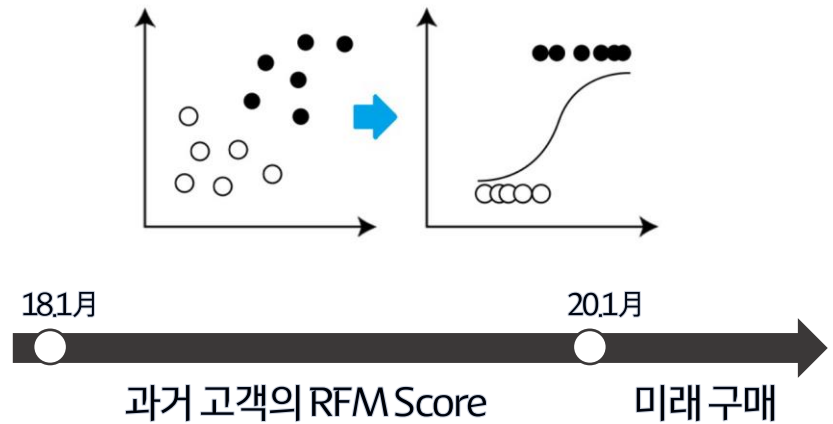
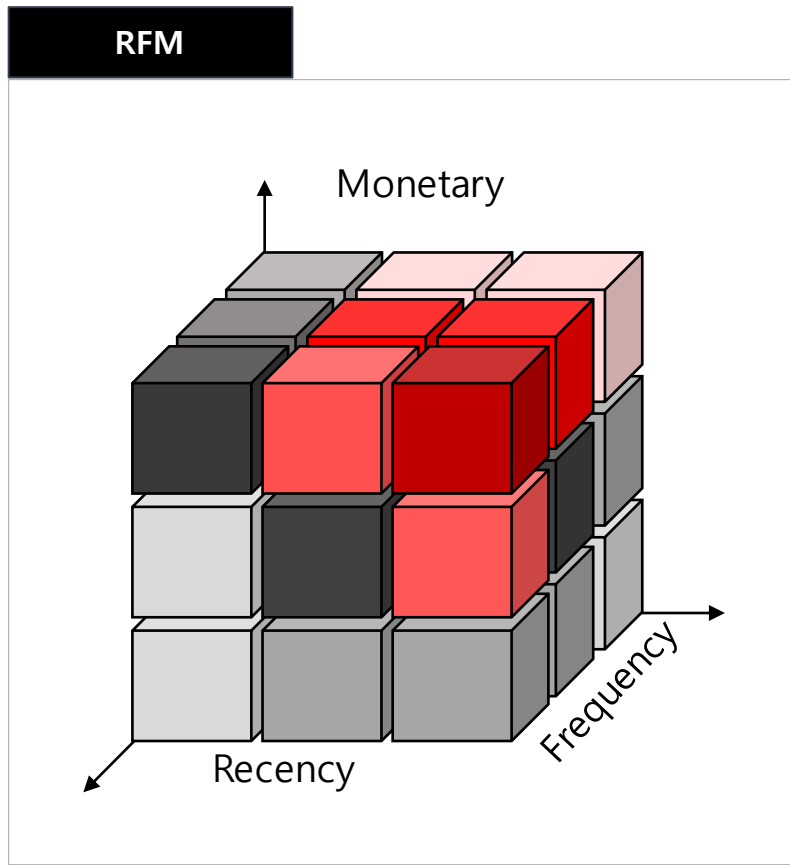
고객 이해를 위한 데이터 모델 설계

고객 Life style 및 행동 style을 정의할 수 있는 변수 개발 “고객 다각적 이해” 추구



4. Data analysis in Marketing

고객 이해를 위한 데이터 모델 설계

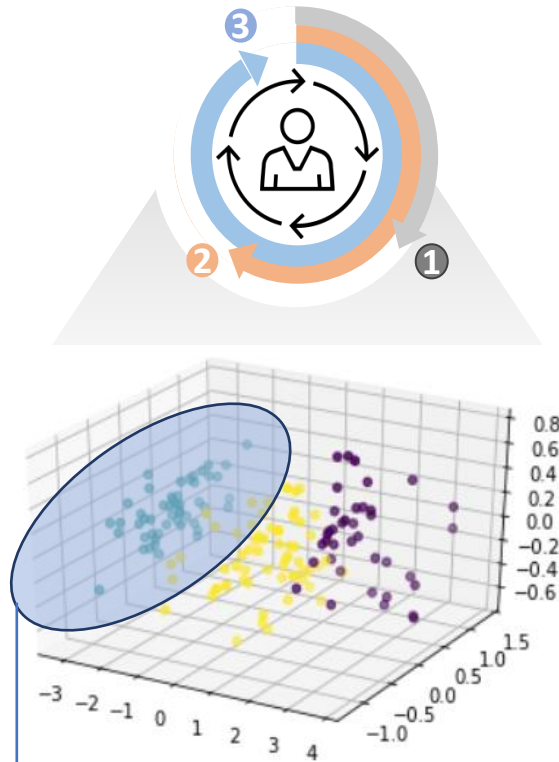


HIGH	높은 가치 고객을 유지하기 위한 전략
MIDDLE	우수 고객으로 전환 하기 위한 가치 상승 전략
LOW	가치 낮은 고객 대상 과도한 서비스 지양 전략

4. Data analysis in Marketing

고객 이해를 위한 데이터 모델 설계

고객 Segmentation



Young 국내뷰티 관심고객군

복잡하고 다양한 고객 속성으로 인한 고객 이해 어려움
고객 Clustering을 통한 명확한 이해기반 “전략 도출”

고객 Cluster별

전시/노출 콘텐츠 차별화, 마케팅 및 제공혜택 차별화

Macro Trend와 Micro Trend를 조합한 남/녀 각 9개 Trend Code



모델 개발 시 “고객 Feature로 활용”

고객 Cluster별 마케팅 목적에 따른 예측 모델 개발

4. Data analysis in Marketing

고객 이해를 위한 데이터 모델 설계

tf-idf selects informative terms

DC-9 WITH 55 ABOARD CRASHES; AT LEAST 16 DEAD
CHARLOTTE, NC, (Reuter)
A USAir DC-9 with 55 people on board crashed and burst into flames during a thunderstorm after missing an approach to Charlotte's international airport Saturday, killing at least 16 people. The flight, which originated in Columbia, South Carolina and was on its final approach, hit a house near the airport runway and caught fire, said Jerry Orr, aviation director at Charlotte-Douglas International Airport. Orr said 16 people were dead, six were missing and presumed dead and 33 were taken to local hospitals. USAir reported 18 dead. Rescue teams fought to save lives inside the wreckage of the plane, which split into three sections on impact at about 6:50 p.m. EDT as the plane was trying to land at Charlotte during heavy storms.

top 15 terms ranked by

frequency	highest idf	tf * idf
32 the	1.00 tdt000077	3.20 orr
16 were	1.00 picknickers	2.81 charlotte
14 said	0.93 screaming	2.65 payne
12 and	0.93 timmy	2.48 dc
12 to	0.86 6thld	2.24 usair
11 a	0.80 orr	2.00 plane
10 of	0.78 1016	1.93 crash
9 at	0.76 bergen	1.74 bones
9 was	0.75 dripping	1.63 survivors
7 in	0.73 abrams	1.50 dripping
6 on	0.72 0419	1.49 wreckage
6 they	0.69 fuselage	1.35 dead
6 people	0.66 nc	1.29 hospitals
6 had	0.66 thunderstorm	1.27 airport
6 plane	0.66 payne	1.23 55

Copyright © Victor Lavrenko, 2014

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

텍스트 분석에서 많이 활용되는 방법



각 문서별 단어 중요도 산출

문서 내용 분별력 높은 단어들에 높은 Score 부여

고객별 주번호 상품 파악을 위해 활용



최근1년 상품카테고리별 구매건수

화장품 스킨	화장품 립스틱	화장품 향수	패션 명품가방	패션 아동복
5	3	1	1	4
TF-IDF				
4.2	1.7	0.8	1.6	4.8

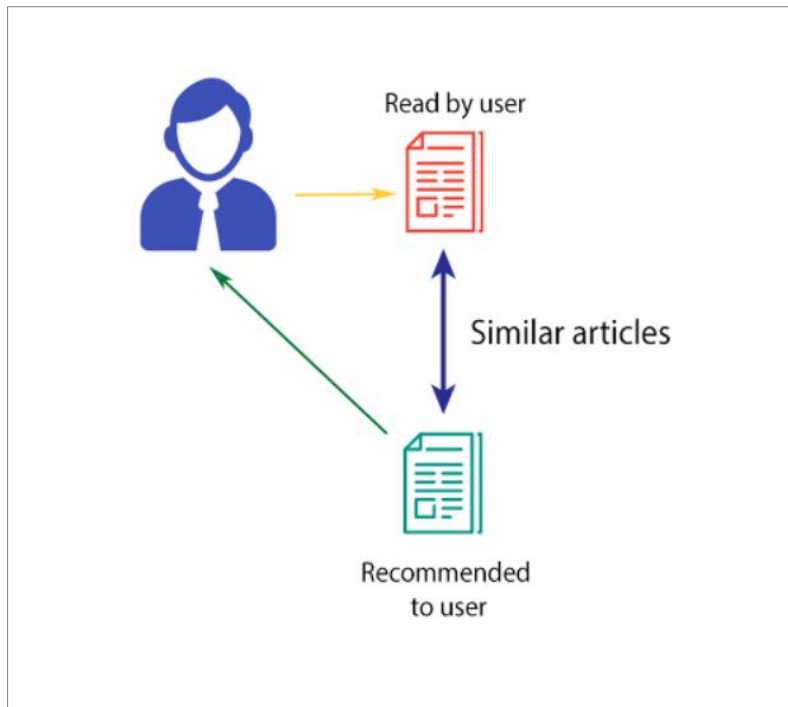
5. 개인화 추천 시스템



5. 개인화 추천 시스템

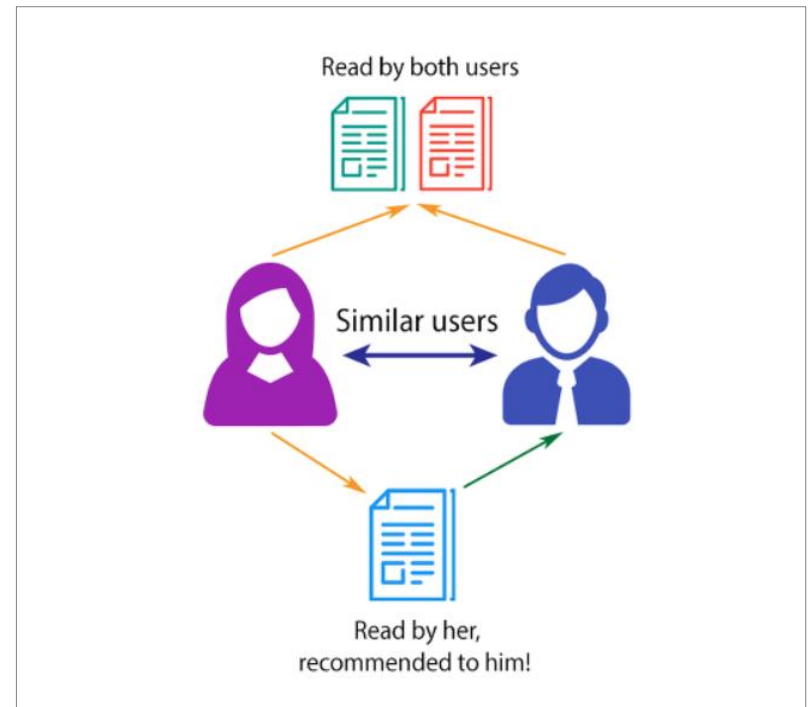
추천시스템 기본 유형

Content based filtering 방법



고객이 특정 아이템을 선호하는 경우
해당아이템의 특성정보를 활용하여
유사한 다른 아이템을 추천하는 방법

Collaborative filtering 방법

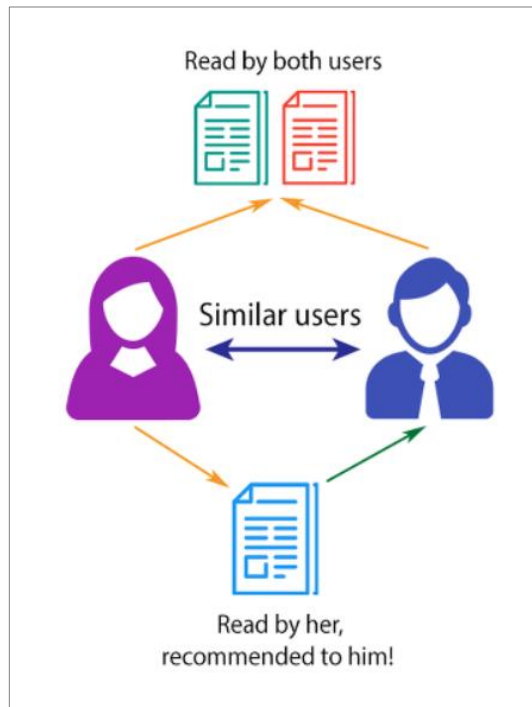


고객과 상품 간의 관계에 대한 분석을 토대로
고객이 선택하지 않은 상품에 대한 추천을
효과적으로 제공하는 방법

5. 개인화 추천 시스템

추천시스템 기본 유형

Collaborative filtering 방법



Model based

“KNN, Matrix Factorization”

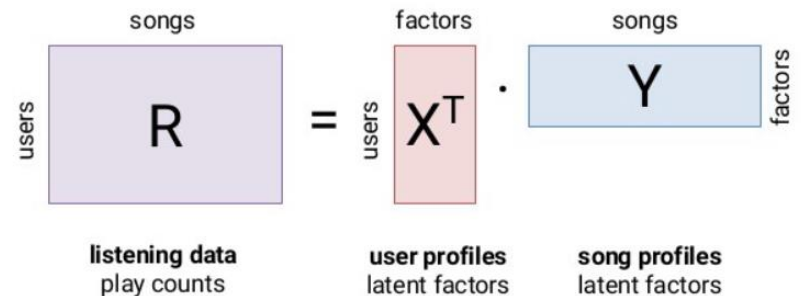
Matrix Factorization

모델
개요

행렬 분해를 통해 분해하고 이를 다시 곱함으로써 기존 행렬 값과 유사하면서 동일한 크기를 가진 행렬을 생성함을 통해 잠재된 의미를 파악하고 이를 활용하여 예측

모델
구조

Model listening data as a product of latent factors

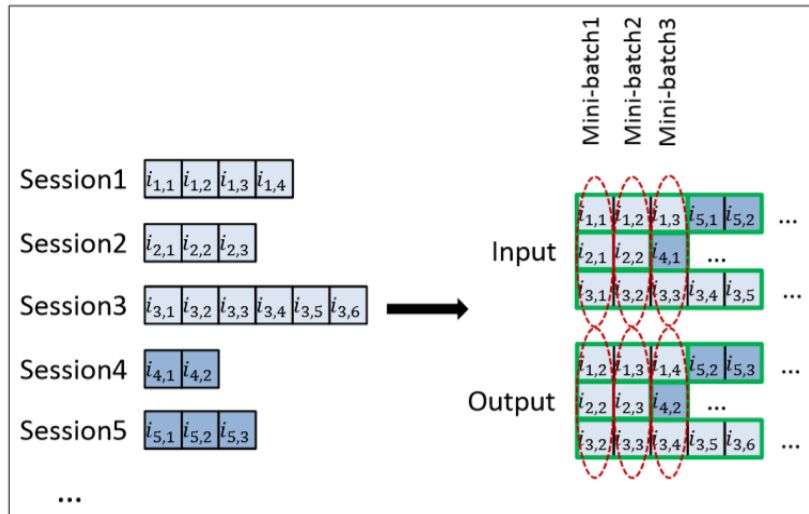


학습
방법

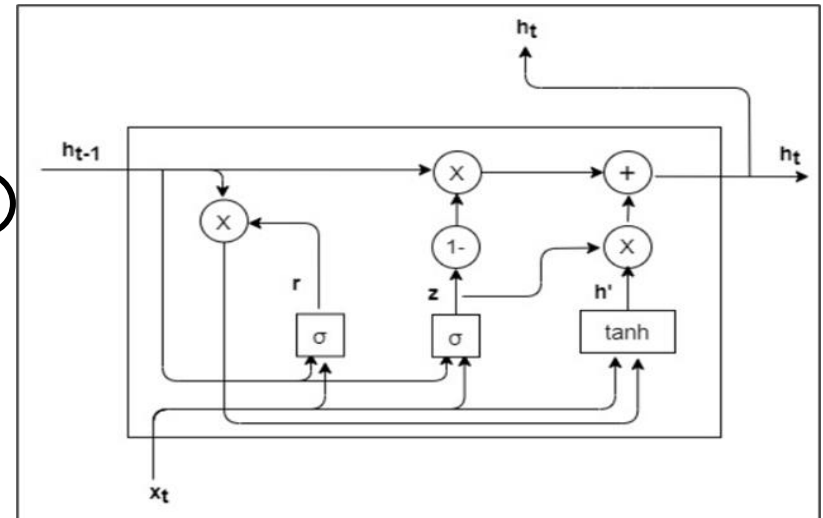
$$\hat{y}(x) := \underbrace{w_0}_{\text{Global bias}} + \sum_{i=1}^n \underbrace{w_i x_i}_{\text{Feature의 Bias}} + \sum_{i=1}^n \sum_{j=1}^n \underbrace{\langle v_i, v_j \rangle}_{\substack{\text{MF 활용하여 Latent} \\ \text{Vector 도출}}} x_i x_j \rightarrow \text{Feature간 Interaction}$$

5. 개인화 추천 시스템

추천모델 + 딥러닝(순환신경망)



Gated Recurrent Unit (GRU)



$$\text{식 (2-11): } z = \sigma(x_t W_x^{(z)} + h_{t-1} W_h^{(z)} + b^{(z)}) \leftarrow$$

$$\text{식 (2-12): } r = \sigma(x_t W_x^{(r)} + h_{t-1} W_h^{(r)} + b^{(r)}) \leftarrow$$

$$\text{식 (2-13): } h' = \tanh(W_x x_t + (r \odot h_{t-1}) W_h + b) \leftarrow$$

$$\text{식 (2-14): } h_t = (1 - z) \odot h_{t-1} + z \odot h' \leftarrow$$

5. 개인화 추천 시스템

추천 모델 평가 : Personalized Ranking

Explicit Feedback	고객이 자신의 선호도를 직접 표현한 Data (리뷰, 평점, 구독 등)
Implicit Feedback	고객이 간접적으로 선호, 취향을 나타내는 데이터 (검색기록, 방문 페이지, 구매 내역 등)
Implicit Feedback 을 이용한 (ex 클릭, 구매) 아이템 추천 모델	<p>관찰되지 않은 고객-아이템 Pair = Negative 피드백 + 결측치</p> <ul style="list-style-type: none">• Negative 피드백 : 고객이 아이템 구매에 관심이 없음• 결측치 : 현재 고객이 선택하지 않았으나, 관심이 있을 수 있음

고객이 $item_i$ 를 선택했을 때 고객이 조회하지 않은 모든 $item_j$ 보다 더 선호함을 가정

선택한 아이템의 r_i 가 선택하지 않은 아이템의 가장 큰 값보다 ($r_j = r_{max}$) 큰 값을 갖는 방향으로 학습

$$P(r_i > r_{max}) = \sum_{j=1}^{N_S} P(r_i > r_j | r_j = r_{max}) P(r_j = r_{max})$$

5. 개인화 추천 시스템

추천 모델 평가 : Personalized Ranking

1. BPR (Bayesian personalized ranking)

베이지안 개인화 순위로 Negative sample (고객이 선택한 상품 외, 상품 중 학습을 위해 샘플링한 상품들 집합) 보다 높은 학습 결과값을 갖는 방향으로 학습되도록 유도

$$L_{bpr} = -\frac{1}{N_S} \cdot \sum_{j=1}^{N_S} \log \sigma(r_i - r_j)$$

2. TOP1

안정적인 학습을 위해 BPR에 $\sigma(r_j^2)$ 규제값 추가

$$L_{TOP1} = \frac{1}{N_S} \cdot \sum_{j=1}^{N_S} \sigma(r_j - r_i) + \sigma(r_j^2)$$

5. 개인화 추천 시스템

추천 모델 평가 : Personalized Ranking

3. BPR - max , TOP1 - max

주어진 Negative sample 이 고객이 선택한 상품의 학습 결과값 대비 상당히 낮은 값들로 이루어진 경우 $(r_i > r_j)$

그러한 다수의 Negative sample 들에 의해 $\sigma(r_j - r_i) \approx 0$ 값을 갖게 되어 Vanishing Gradient 문제 발생

손실함수에 $\text{Softmax}(r_j)$ 값을 추가하여

Score가 높은 Negative sample 위주로 학습할 수 있도록 변경 함으로써 Vanishing Gradient 문제 해결

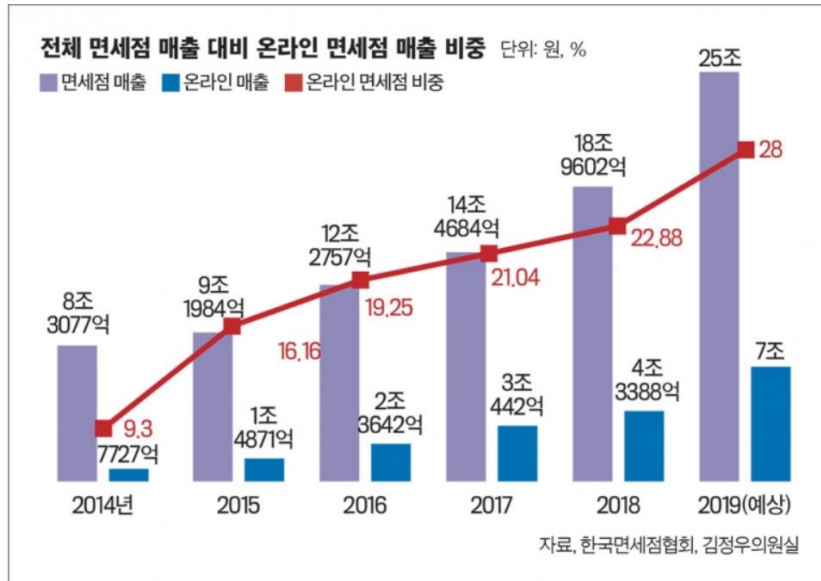
- Target score가 낮을 때, 전반적으로 parameter값 업데이트됨
- Target score가 높을수록 max negative sample에 집중해서 parameter값을 업데이트 함

$$L_{TOP1_max} = \sum_{j=1}^{N_s} s_j (\sigma(r_j - r_i) + \sigma r_j^2)$$

$$L_{bpr_max} = -\log \sum_{j=1}^{N_s} s_j (\sigma(r_i - r_j))$$

5. 개인화 추천 시스템

추천 시스템 연구 배경



- 최근 5년 온라인 시장에서 고객 구매 비중 급속도로 증가
- 가격 경쟁 중심에서 고객 경험 개선 중심으로 변화
- 서비스 측면 차별화 및 경쟁 우위 확보 중요
- 해외 여행 제약을 가지고 있는 면세 특성 상 효율적 마케팅 必

고객 취향 및 구매 패턴에 대한 이해를 바탕으로 고객 경험을 제고할 수 있는 **개인화 추천 서비스 도입에 대한 관심 증대**

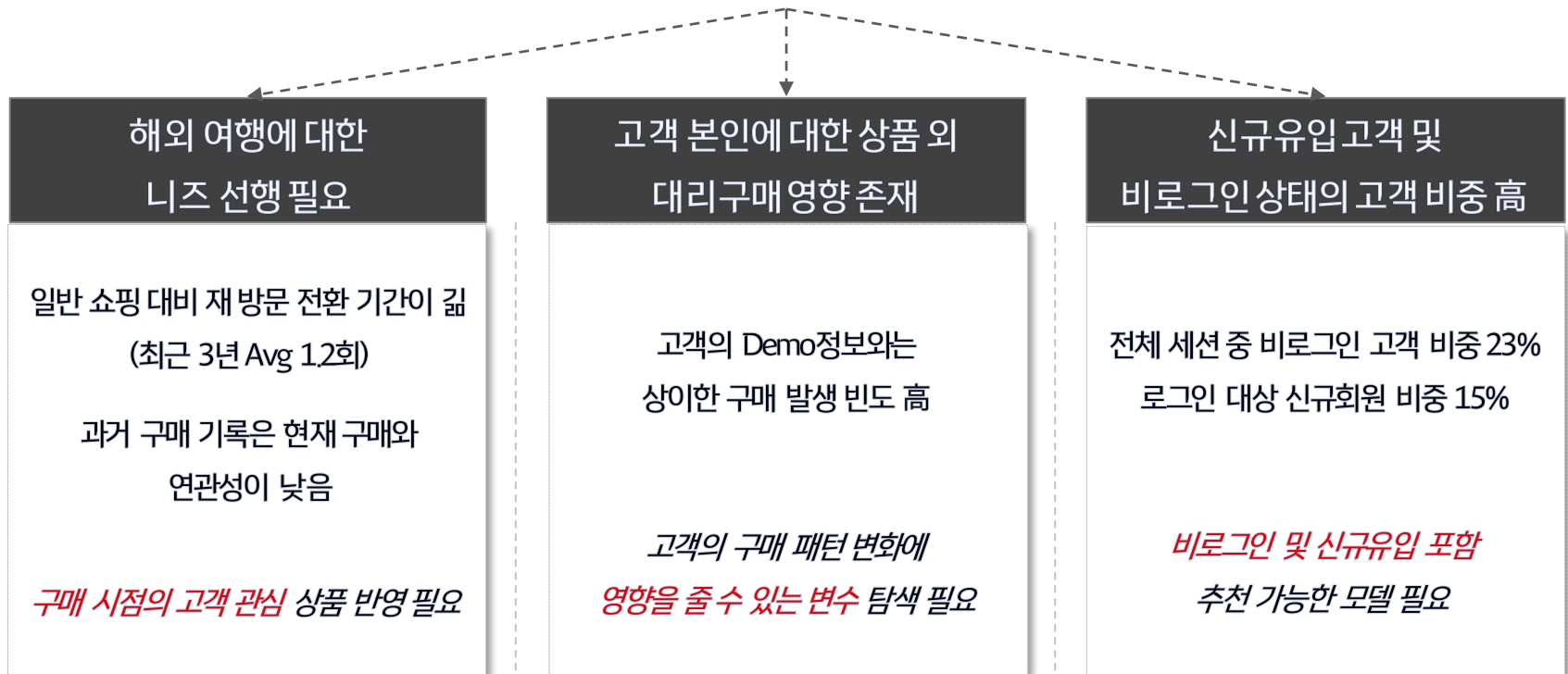
현재 일반 쇼핑몰 데이터를 활용한 추천 서비스에 대한 연구는 많이 진행되었으나 **면세점 데이터를 활용한 연구는 부재함**

면세점의 실 데이터를 활용하여 면세점에 최적화된 추천 모델에 대한 연구 필요

5. 개인화 추천 시스템

추천 모델 기본 방향

면세점 데이터의 경우 일반 쇼핑 플랫폼과 다른 특성을 지님



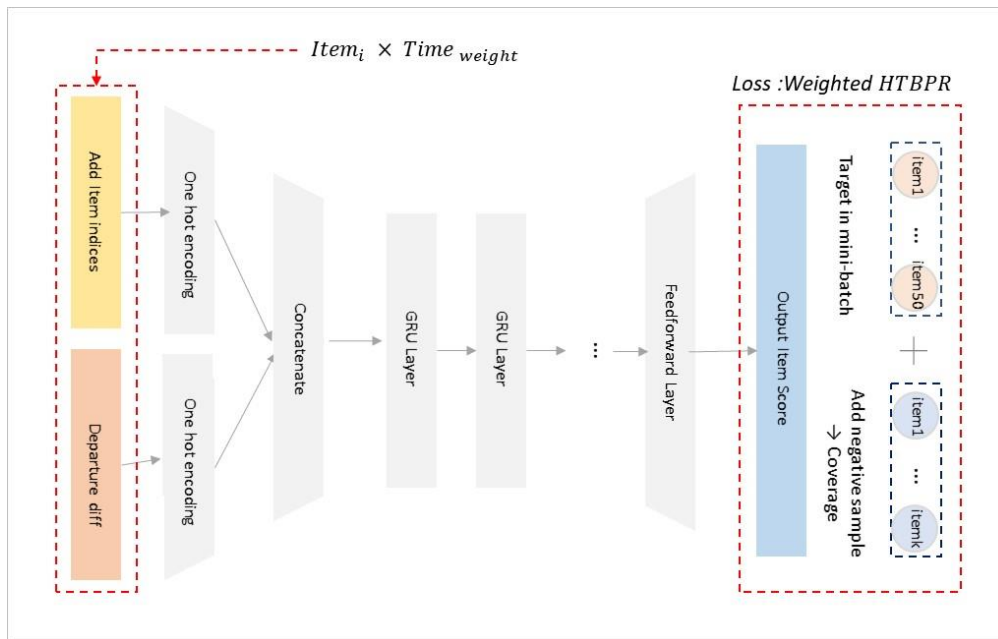
Session 기반의 Sequential 한 상품 클릭 정보를 기반으로 한 추천 모델 구현 진행

5. 개인화 추천 시스템

면세점 특성을 반영한 모델 구조

면세점 데이터의 특성을 반영한 모델 최적화를 통해

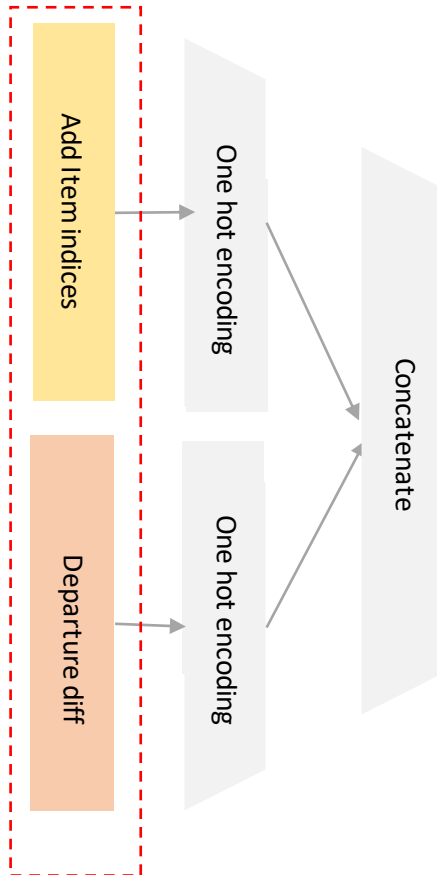
기존 추천 모델 성능 대비 MRR@20 **21%**, Recall@20 **15%** 성능 향상



- 면세점 고유 특성 반영 위한 출국도래일 정보 추가
- 각 상품의 클릭 간 체류 시간 정보에 따라 가중치가 부여되도록 학습 데이터 재구성
- 면세점 데이터에 최적화된 손실 함수 제안 (Weighted Hyperbolic Tangent Based BPR)
- Coverage 기반 샘플링 상품 Negative sample로 활용

5. 개인화 추천 시스템

모델 학습을 위한 변수 처리



- 모델 학습을 위한 변수 처리

Input변수로 활용되는 텍스트 정보에 대해 원-핫 인코딩을 통한 벡터 변환

- 면세점 연관 변수 추가

1. 출국도래일

- 1) 면세점 구매의 경우 해외 여행 전제 必

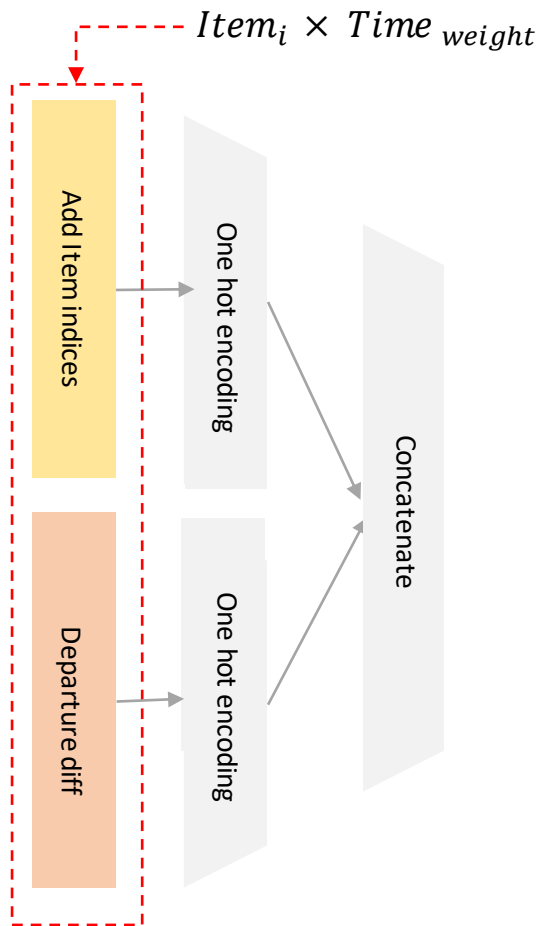
- 2) 여행 출국일 60일 전 ~ 출국 당일까지 한정된 기간 동안 구매 가능

- 3) 구매 상품에 대해 즉시 수령이 아닌 출국당일 인도장에서 구매한 상품에 대해 일괄 수령

→출국도래일에 따라 고객 관심 상품에 변화 발생

5. 개인화 추천 시스템

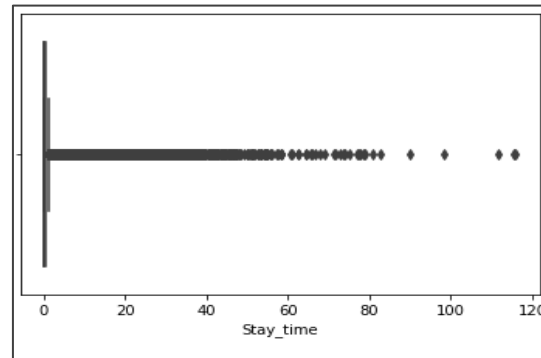
페이지 체류시간 가중치 적용



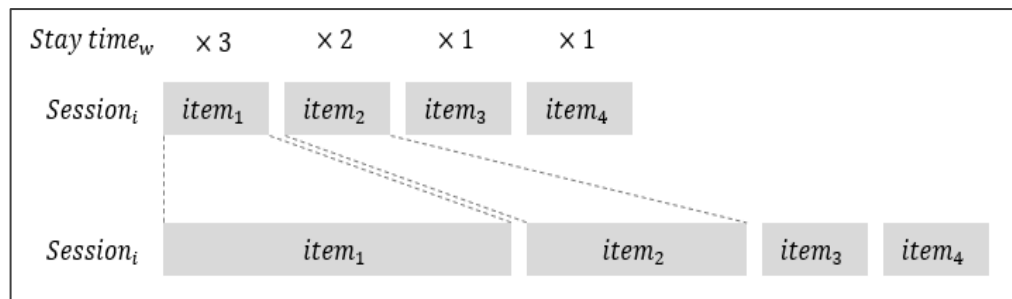
페이지 내 체류 시간의 정도 == 상품에 대한 고객의 관심정도 반영

전체 데이터의 80% 1분 이하의 짧은 체류 시간 보유

4개 구간으로 범주화 한 후 구간별로 다른 Weight 적용



기준	Index	Weight
Mean	0.71	
Std	2.12	
Min	0.00	1
25%	0.10	1
50%	0.22	2
75%	0.53	3
90%	1.25	4
Max	115.92	4

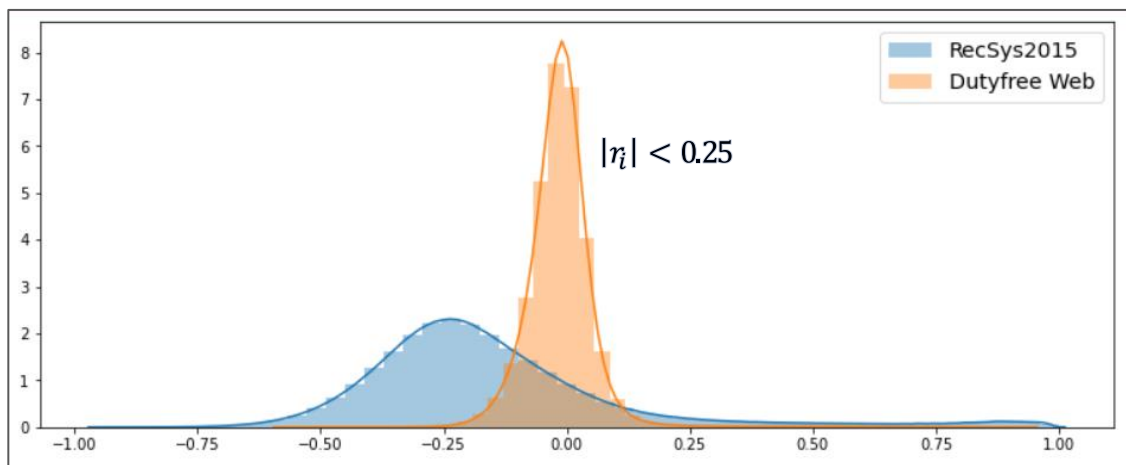


5. 개인화 추천 시스템

손실함수 개선을 통한 모델 최적화

가격적인 측면에서 일반 이커머스 대비 강점을 가지고 있어 개인만을 위한 소비가 아닌 다른 사람을 위한 구매가 빈번하게 발생하고 있음
동일 고객 유형에 대해 **다양한 구매 패턴 발생** 가능

기본 GRU 모델 학습 결과 상품 Score



일반 이커머스 대비

$|r_i| < 0.25$ 인 분포를 보이고 있음

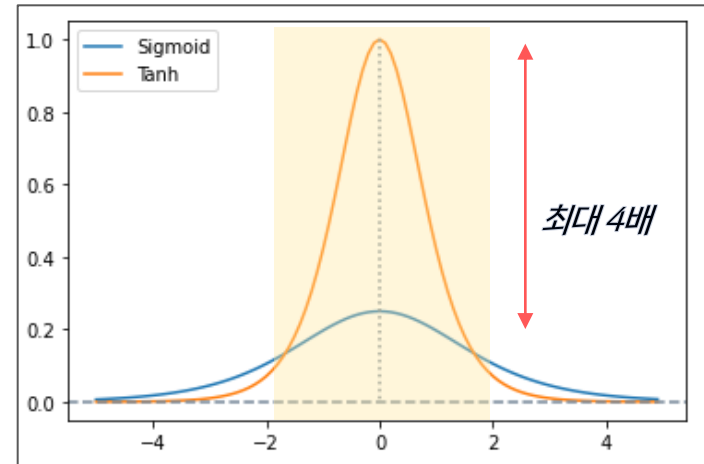
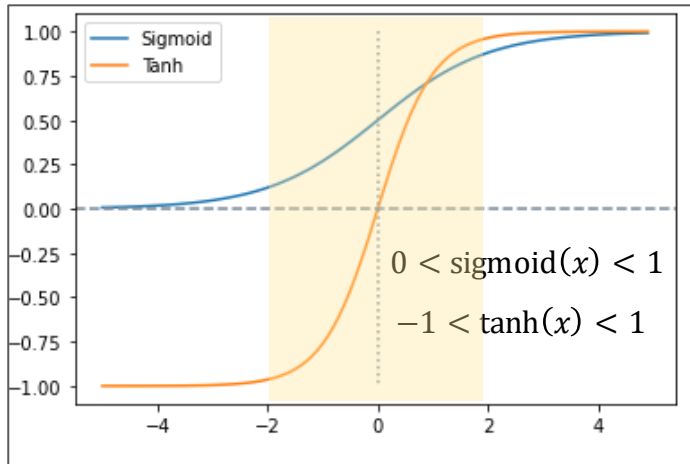
$$L_{\text{top1-max}} = \sum_{j=1}^{N_S} s_j \left(\sigma(r_j - r_i) + \sigma(r_j^2) \right) \quad r_j(\text{Negative}) \text{와 } r_i(\text{Positive}) \text{ 값의 차이 } \sigma(r_j - r_i) \approx 0 \text{ 매우 작은 값이 손실 값으로 계산 됨}$$

유사한 값들의 차이를 잘 구분하고 이를 학습할 수 있는 손실함수 적용 필요

5. 개인화 추천 시스템

손실함수 개선을 통한 모델 최적화

손실함수로 제안 : Weighted Hyperbolic Tangent based BPR (W-HTBPR)



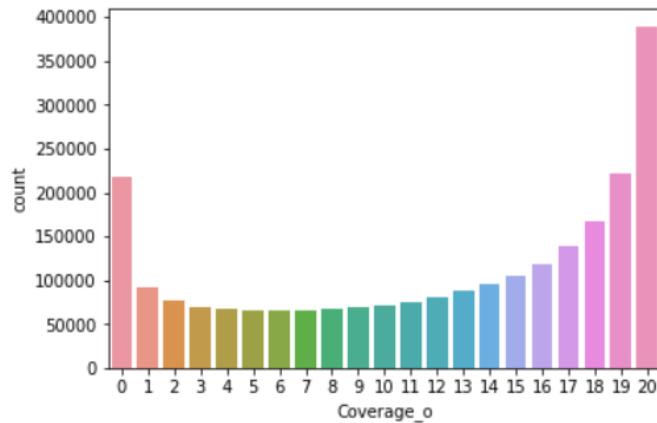
면세 특성을 반영한 다양한 Input 변수 추가 및 Negative sample이 추가된 모델에서 학습된 값의 매우 작은 차이를 잘 학습

$$L_{W-HTBPR} = \sum_{j=1}^{N_s} s_j (\tanh(r_j - r_i) + \tanh(r_j^2))$$

5. 개인화 추천 시스템

Coverage기반 Negative sample 추가

GRU 학습 결과 고객이 클릭한 상품 대비 Coverage 높은 상품 수

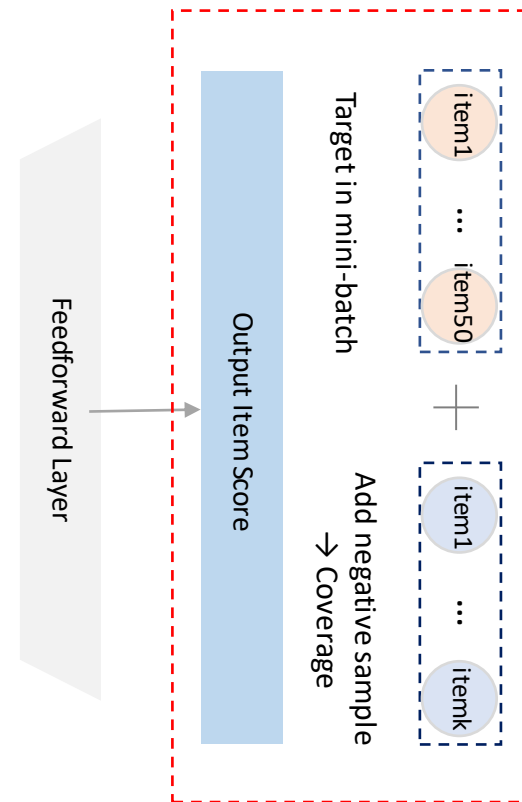


Coverage기반 상품을 샘플링하여
고객이 관심을 가질만한 상품이지만 상품을 인지하지 못한 경우 포함
학습할 수 있도록 Negative sample 추가

Sampling probability : $Coverage^\alpha$

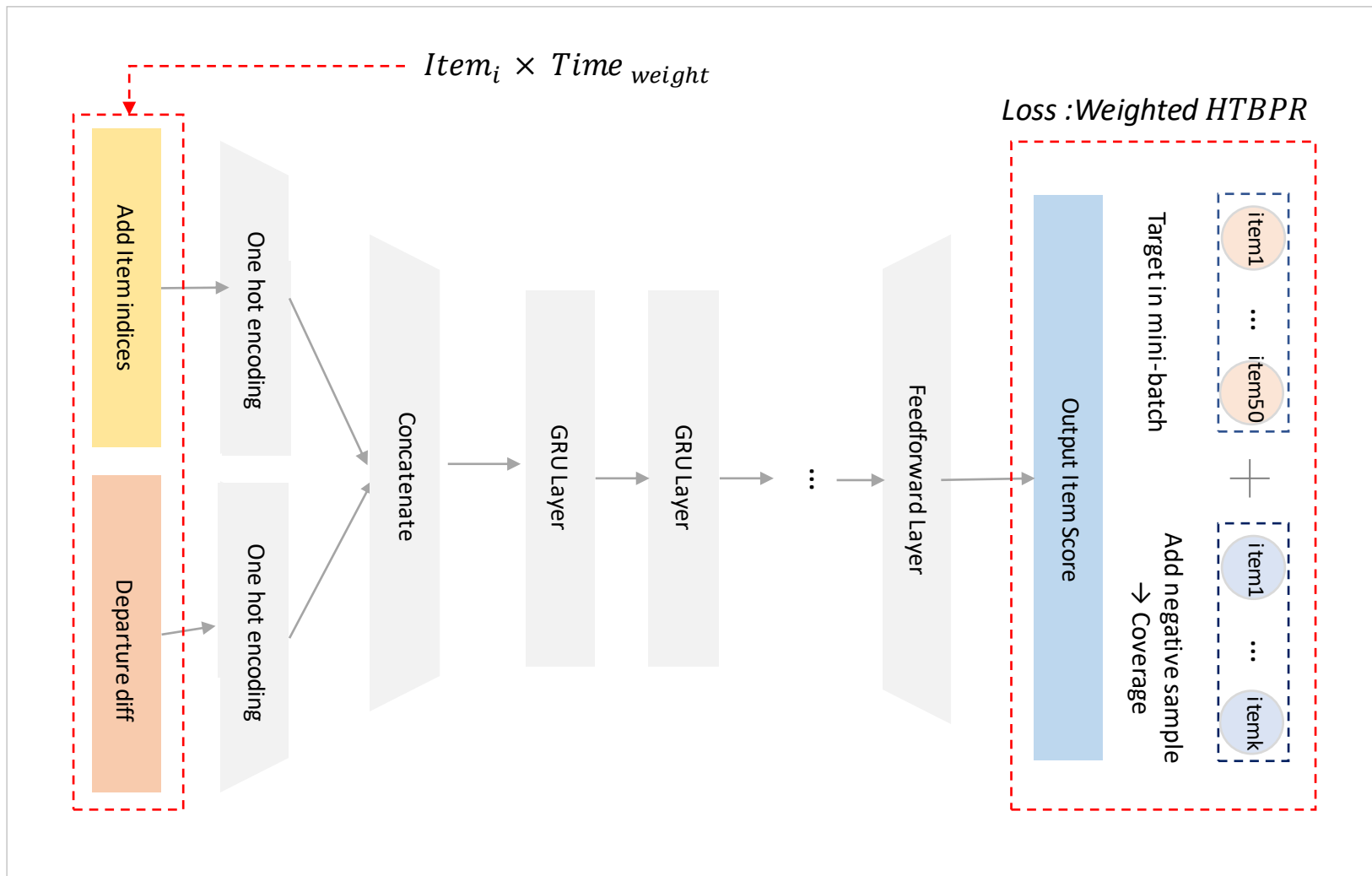
- Alpha = 0 : uniform
- Alpha = 1 : Coverage based

Loss : Weighted HTBPR



5. 개인화 추천 시스템

최종 제안 모델



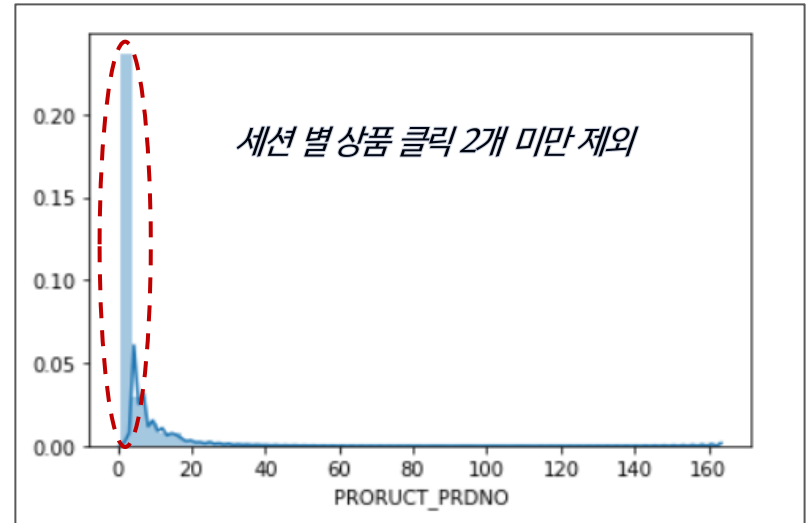
5. 개인화 추천 시스템

데이터 상세 - DutyFree

면세점 WEB (PC/Mobile) 로그 데이터

구분	내용
채널	WEB (PC/MOBILE)
데이터 기간	2019.12.01 ~ 2019.12.31
총 상품 클릭 로그 수	7,865,245
총 세션 수	1,557,259
총 카테고리 수	1,221
총 Item 수	81,982
상품 View time 정보	평균 0.71분
페이지 특성 수	3,261

구분	세션 수	클릭 수
Train dataset (12.01 ~ 12.19)	591,305	4,326,968
Test dataset (12.20 ~ 12.31)	353,076	2,761,042



Train data 데이터 내 존재하지 않는

상품 클릭 데이터 제외

5. 개인화 추천 시스템

데이터 상세 – RecSys Challenge 2015

“A sequence of click events performed by some user during a typical session in an e-commerce website”

면세점과 동일하게 데이터 처리 후 실험 진행

1) 세션 별 상품 클릭 2개 미만 제외

2) Train data 내 존재하지 않는 상품 클릭 데이터 제외

구분	세션 수	클릭 수
Train dataset	7,990,018	31,744,233
Test dataset	1,997,887	7,934,563

5. 개인화 추천 시스템

모델 성능 평가지표

1. Recall@K : 전체 추천 중 추천 결과값의 상위 K내에 적합 상품이 추천된 횟수에 대한 비율

$$Recall@K = \frac{\text{상위 K 내에 적합 상품이 추천된 횟수}}{\text{전체 추천수}}$$

2. MRR@K : 단순 상위 K내에 있는지 여부가 아닌 순위에 대한 결과값도 함께 평가하기 위한 지표

$$MRR@K = \frac{1}{|U|} \sum_{i=1}^{|U|} \frac{1}{r_i}$$

실제 학습된 추천 모델을 기반으로 추천 화면을 구성 시 단순히 사용자에게 적합한 상품인지를 판단하는 것 외에
상위 순위에 노출되는지에 대한 정보는 실제 환경에서 매우 중요

2가지 성능 측정 지표를 모두 고려하여 모델 평가

5. 개인화 추천 시스템

Baseline 실험 결과

모델	Loss	Evaluation Recall			Evaluation MRR		
		20	10	5	20	10	5
ItemKNN		0.231	0.169		0.085	0.081	
MF	BPR	0.036			0.016		
GRU4REC	Cross-Entropy	0.224	0.186	0.150	0.107	0.105	0.100
GRU4REC	TOP1	0.335	0.276	0.219	0.152	0.148	0.141
GRU4REC	BPR	0.203	0.171	0.140	0.101	0.099	0.094
GRU4REC	BPR-max	0.279	0.228	0.181	0.127	0.123	0.117
GRU4REC	TOP1-max	0.360	0.296	0.232	0.157	0.154	0.145
+ 카테고리	TOP1-max	0.041			0.021		

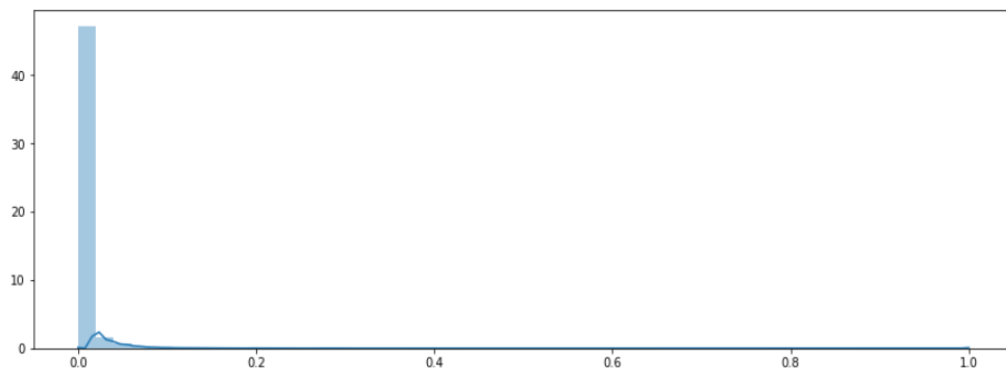
단순 상품간 연관성 정보 및 행렬 분해를 통한 상품 예측은

상품 추천 시스템 모델로 활용 시 성능 향상에 한계 존재

일반 이커머스에서 자주 활용되고 있는

카테고리 정보를 추가할 경우 모델 성능이 급격하게 낮아짐

각 상품의 카테고리 간 Similarity



선물, 대리구매 등 개인 외 다른사람을 위한 구매

= 다양한 구매 패턴 발생

Category 간의 유사성 대부분 0.1 미만으로 매우 낮아짐

Category를 input변수로 활용했을 경우 학습 성능이

현저하게 떨어지는 결과를 가져옴

5. 개인화 추천 시스템

Baseline 실험 결과

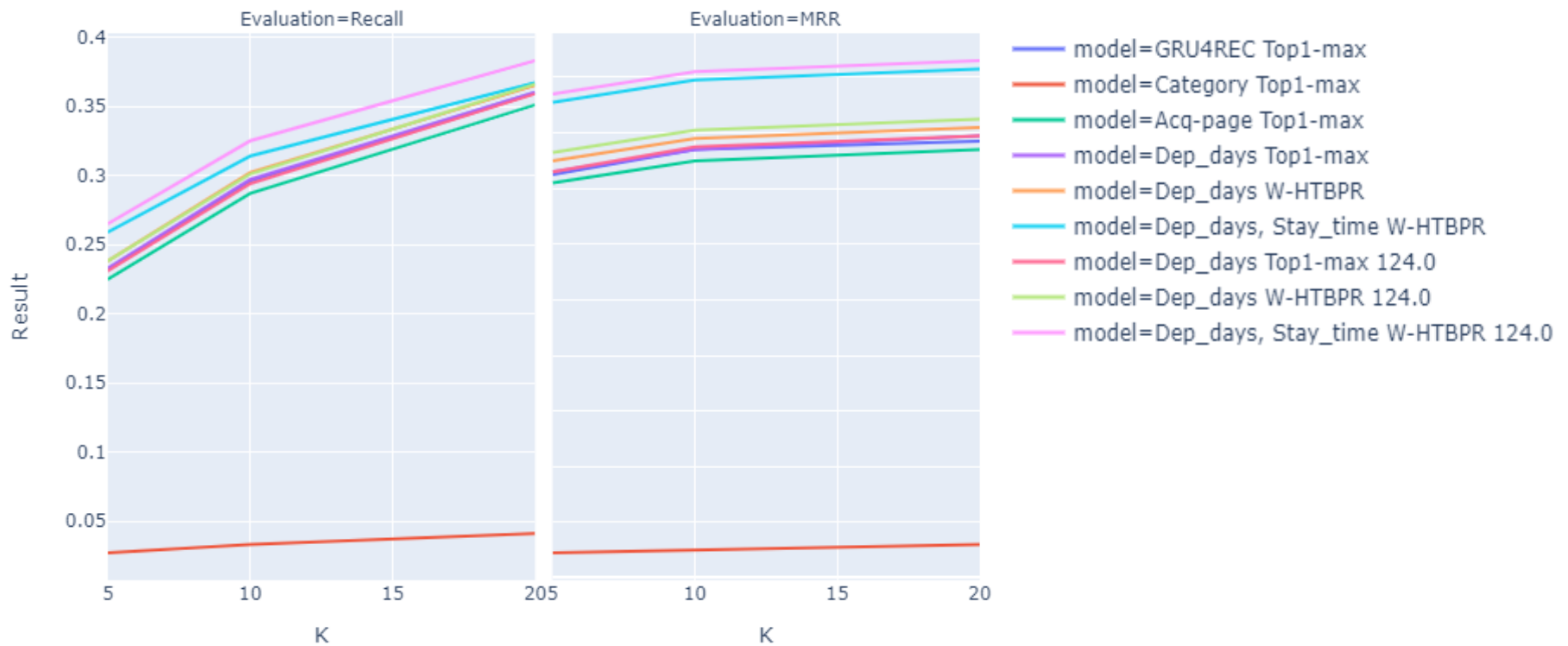
면세점 데이터에는 새롭게 만든 W-HTBPR Loss 함수 적합

모델	Loss	샘플 추가	Evaluation Recall			Evaluation MRR		
			20	10	5	20	10	5
Dutyfree	Top1-max		0.360	0.296	0.232	0.157	0.154	0.145
Dutyfree	W-HTBPR		0.366	0.302	0.238	0.162	0.158	0.149
Dutyfree	Top1-max	1024	0.359	0.293	0.229	0.159	0.155	0.146
Dutyfree	W-HTBPR	1024	0.366	0.302	0.238	0.165	0.161	0.152
RecSys2015	Top1-max		0.586	0.497	0.388	0.255	0.249	0.234
RecSys2015	W-HTBPR		0.586	0.498	0.390	0.256	0.250	0.236
RecSys2015	Top1-max	1024	0.630	0.528	0.402	0.261	0.254	0.237
RecSys2015	W-HTBPR	1024	0.626	0.524	0.398	0.260	0.253	0.236

5. 개인화 추천 시스템

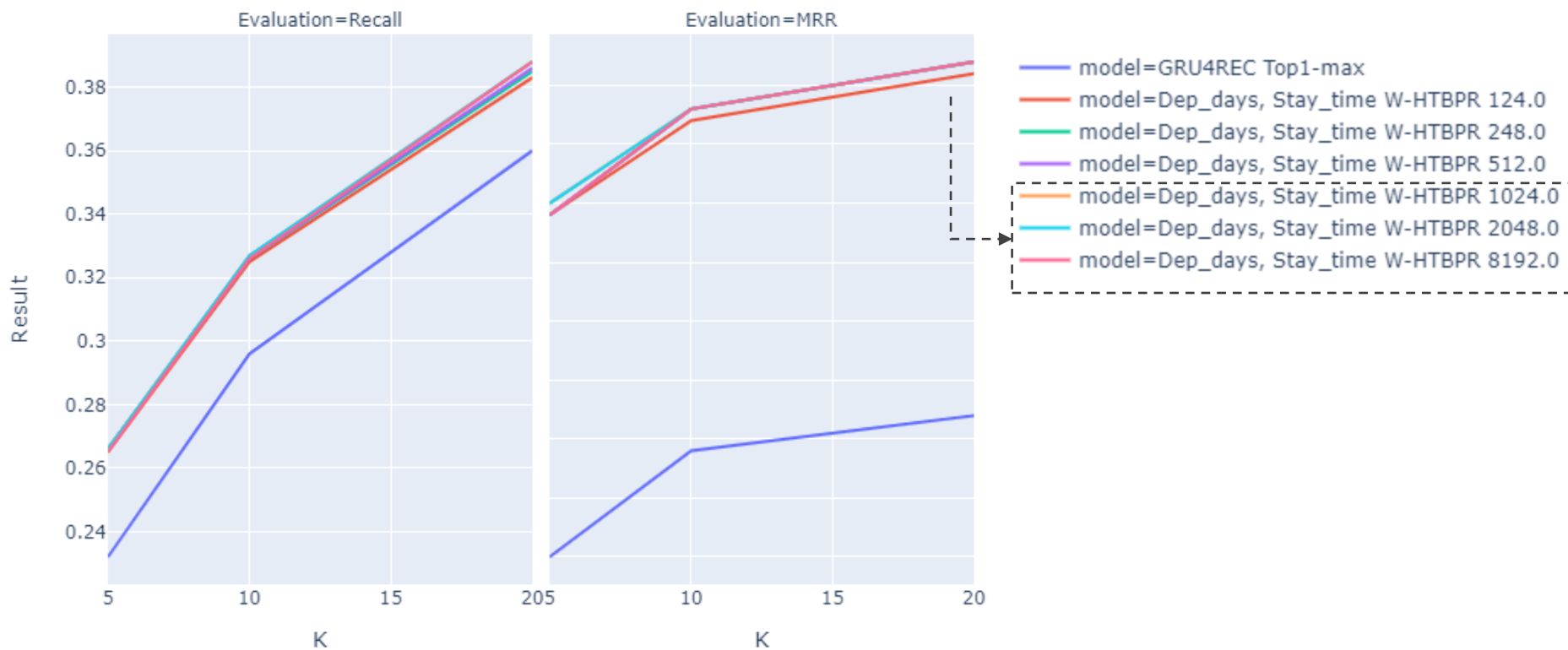
출국도래일,체류시간가중치 적용,W-HTBPR,Add sample 124에서

기존 모델 대비 MRR 20%,Recall 14% 향상 됨



5. 개인화 추천 시스템

Add sample 1024에서 기존 모델 대비 MRR 21%, Recall 15% 향상 됨



5. 개인화 추천 시스템

엠엘비 MLB
CPHE 웨도우 버킷햇 뉴욕 양키스 BLACK

MLB



INPUT

프레쉬 FRESH
로터스 프리저브 레스큐 마스크 100ml

fresh



TARGET

한글 KANGOL
코튼 버킷햇 2117 네이비

KANGOL



RANK1

엠엘비 MLB
CPHE 웨도우 버킷햇 BEIGE 59

MLB



RANK2

프레쉬 FRESH
로터스 프리저브 레스큐 마스크 100ml

fresh



RANK3

디어달리아 DEAR DAHLIA
파라다이스 드림 벨벳 립 무스 - 구아바

DEAR DAHLIA



RANK4

베리맘 VERYMOM
베리맘 시카 세라마이드 크림 80g

VERYMOM



RANK5

배럴 BARREL
배럴 서프 버킷햇 V3 블랙

BARREL



RANK6

감사합니다