

Chapter 39.[실습39] Ecommerce 고객 상품 구매 예측

1 데이터 소개와 분석 개요

데이터 출처와 요약 설명

1.

데이터 소개와
분석 개요

Ecommerce 상품 구매 데이터

- 출처 : <https://rees46.com/en/datasets>
- 데이터 요약 설명 : **Ecommerce** 플랫폼에서 상품에 반응한 유저의 데이터
- 데이터 수 : 67501979 (본 실습에서는 10프로만 sampling하여 사용)
- 테이블 수 : 1
- 데이터 컬럼 수 : 9

데이터 설명 - Table 단위

1.

데이터 소개와
분석 개요

No.	변수명	설명	비고
1	event_time	event가 일어난 시간	UTC
2	event_type	Ecommerce에서 특정 상품에 대한 사용자의 행동	3가지 Type view cart purchase
3	product_id	상품 ID	
4	category_id	상품 카테고리 ID	
5	category_code	상품 카테고리 코드	
6	brand	상품 브랜드	
7	price	상품 가격	
8	user_id	user id	
9	user_session	user session	

데이터 분석 목적

1.

데이터 소개와
분석 개요

주요 목적

- **E-Commerce** 사용자들의 데이터를 분석하여, 사용자들의 구매 패턴에 따른 군집화를 진행
- 군집화 결과를 활용해서 추후 마케팅, 추천에 활용

데이터 분석 목적

1.

데이터 소개와
분석 개요

분석 프로세스

Loop



- | | | | |
|-------------|-------------------|----------------|--------------------|
| - 데이터 수집 | - Data Join | - Data Summary | - Base Model |
| - 분석 문제점 정의 | - Duplicated Data | - Data Explore | - Optimization |
| | - Data Validation | - Insight | - Model Evaluation |
| | - Missing Value | | - Model Analysis |
| | - Normalization | | - 추가 Analysis |

데이터 분석에서 완벽히 순차적으로 분석할 수는 없다!
순서가 조금씩 바뀌거나 여러 작업이 반복되기도 하며 어떤 작업은 더 선행되어 진행될 수 있다!

데이터 분석 목적

1.

데이터 소개와
분석 개요

분석 머신러닝 모델

- Clustering Algorithm 활용
- K-means Algorithm

Clustering 알고리즘 리뷰

- 클러스터링(Clustering)은 비지도학습(Unsupervised Learning)의 대표적인 알고리즘.
- 데이터를 비슷한 특성을 가진 그룹으로 나누는 작업.

1. K-Means 클러스터링 알고리즘

2. **K-Means** 알고리즘은 가장 대표적인 클러스터링 알고리즘 중 하나. 주어진 데이터를 K 개의 클러스터로 나누는 알고리즘으로, 클러스터 중심값을 초기값으로 설정하고 모든 데이터 포인트를 가장 가까운 클러스터에 할당한 뒤, 할당된 데이터 포인트들을 이용해 클러스터 중심값을 재계산. 이 과정을 반복하여 클러스터 할당과 중심값 업데이트를 진행.

Clustering 알고리즘 리뷰

2. 계층적 클러스터링 알고리즘

1. 계층적 클러스터링 알고리즘은 거리 기반으로 클러스터를 형성하는 알고리즘. 데이터 포인트들을 하나씩 묶어가며 계층적으로 클러스터를 형성. 이 알고리즘은 Bottom-Up 방식인 Agglomerative Clustering과 Top-Down 방식인 Divisive Clustering으로 나눌 수 있음.

Clustering 알고리즘 리뷰

3. 밀도 기반 클러스터링 알고리즘

1. 밀도 기반 클러스터링 알고리즘은 데이터 포인트들이 서로 밀도가 높은 영역에 뭉쳐있는 경우 클러스터를 형성하는 알고리즘.

DBSCAN(Density-Based Spatial Clustering of Applications with Noise)과 OPTICS(Ordering Points To Identify the Clustering Structure)가 대표적인 알고리즘.

4. 모델 기반 클러스터링 알고리즘

1. 모델 기반 클러스터링 알고리즘은 데이터가 특정 확률 모델에 따라 생성되었다는 가정을 기반으로 클러스터를 형성하는 알고리즘. **Gaussian Mixture Model(GMM)**과 **Latent Dirichlet Allocation(LDA)**가 대표적인 알고리즘.

Clustering 알고리즘 리뷰

- 클러스터링 알고리즘은 다양한 종류가 있으며, 각각의 알고리즘이 데이터의 특성에 따라 적합한 결과를 제공. 선택한 알고리즘에 따라 클러스터링 결과가 다르게 나타날 수 있으므로, 사용 목적과 데이터 특성에 맞는 알고리즘을 사용해야함.

Key Point

1.

데이터 소개와
분석 개요

본 실습에서 얻어갈 수 있는 Key Point

- Sparse 데이터를 처리하는 방법
- Unsupervised Learning 기반의 데이터 분석 방법
- K-means Model 학습과 최적화 방법

Chapter 39.[실습39] Ecommerce 고객 상품 구매 예측

2 데이터 전처리 1

Chapter 39.[실습39] Ecommerce 고객 상품 구매 예측

3 데이터 탐색

Chapter 39.[실습39] Ecommerce 고객 상품 구매 예측

4 데이터 전처리 2

Chapter 39. [실습39] Ecommerce 고객 상품 구매 예측

5 머신러닝 모델기반 데이터 분석

Chapter 39.[실습39] Ecommerce 고객 상품 구매 예측

6 문제 해결 insight

문제 해결 Insight

6.

문제 해결
Insight

Insight

- 너무 **sparse** 데이터 분포를 띄면, 이를 최소화할 수 있는 데이터 가공법을 고려한다.
 - **category** 데이터를 활용하여 **sparse**함을 줄인다.
- 타겟(**label**)이 없는 경우는 클러스터링을 활용한다.
- **K-means** 알고리즘은 **k** 수를 설정해줘야 한다. 이를 **optimization**할 수 있다.