『Marketing Science』

Week 8 : Segmentation and RFM

조보연 | Gmarket

목차

Causal-Lab

- Understanding the source of Customer Dynamics
- Analyzing customer dynamics - Segmentation
- Analyzing customer dynamics - RFM

Customer Dynamics

Causal-Lab

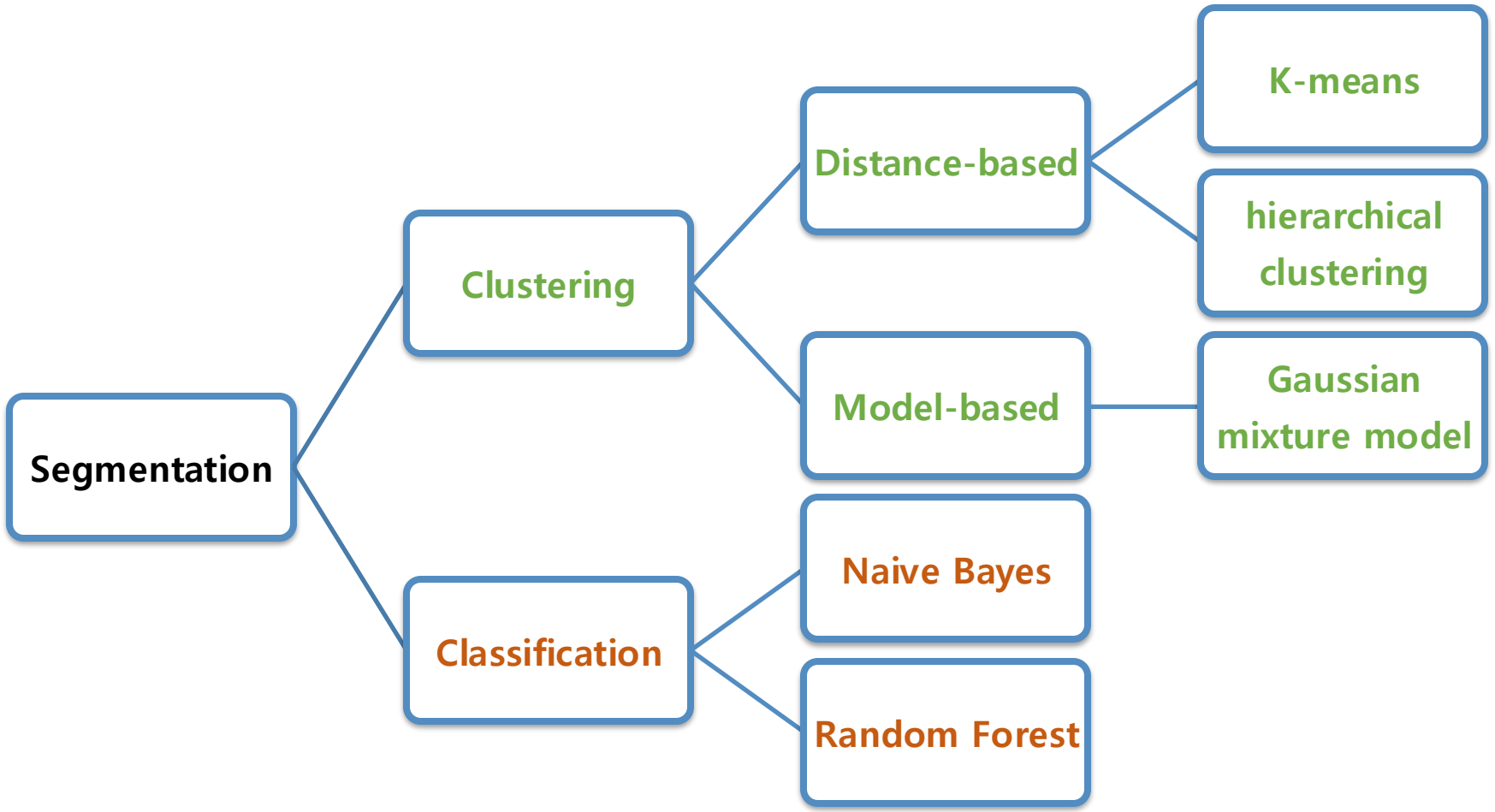
전반적인 고객 생애주기에 대해 이해할 수 있는 이벤트이며, 고객 세분화에 많은 도움을 제공함

Event	Rate of change	Level	Examples
Discrete life events	Rapid	Individual	Parents develop a new purchase pattern, such as gyms that offer daycare
Typical life cycle, maturation	Slow	Individual	Older customers tend to seek reduced risk
Product learning effects	Medium	Individual	By using a product, customers identify additional high-tech features they would like
Product life cycle	Medium	Product market	Initially, users may pay more for new features, before they become price-sensitive
Changes in economy, govt, industry, culture	All	Environment	Recommendations and preferences for healthy food

Table 7.1 – Categories of customer dynamic sources

Segmentation

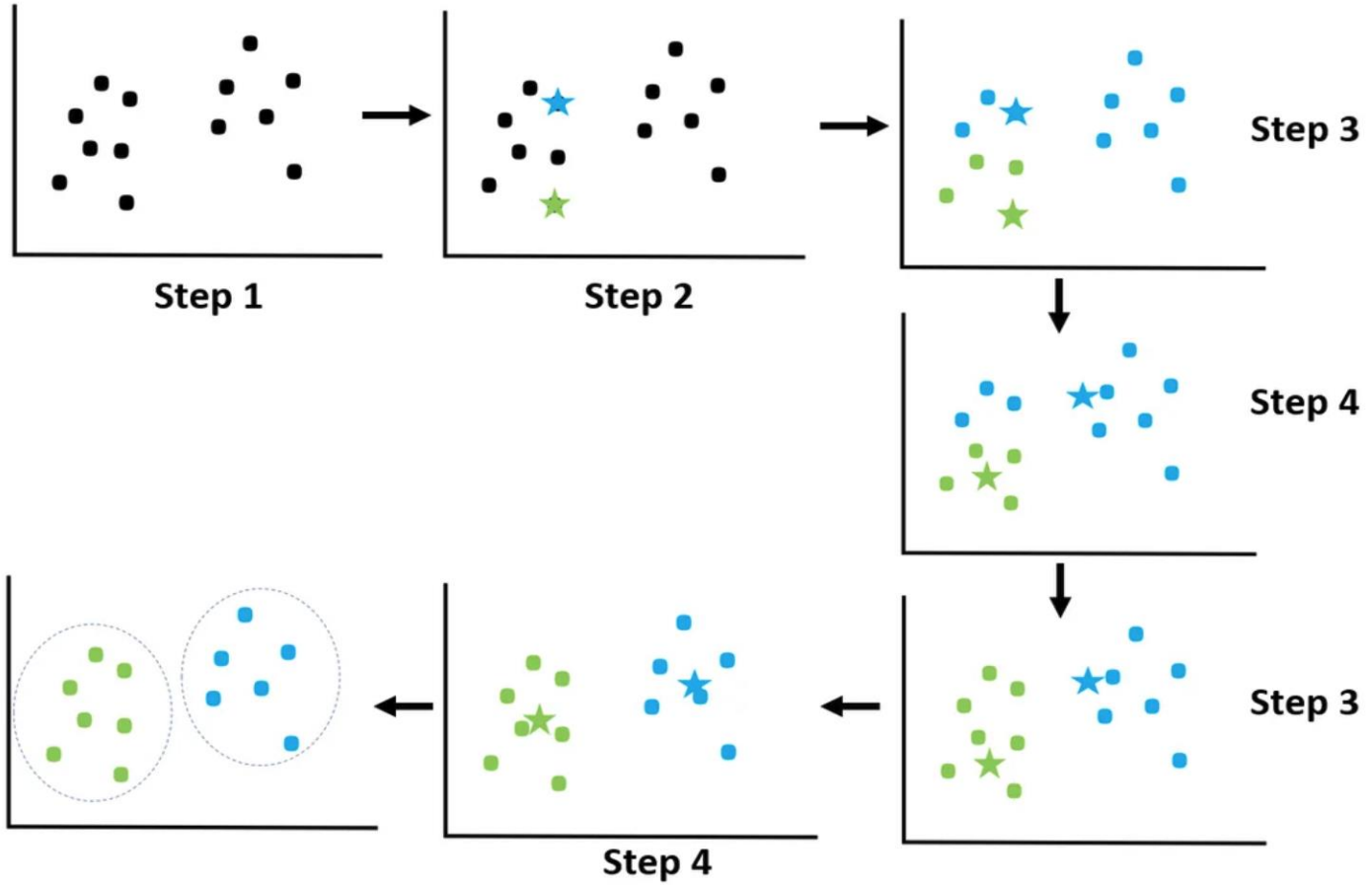
공통적인 특성(니즈/선호)을 기준으로 고객들을 그룹으로 나눈 것이며,
정답셋 유무에 따라 분석 기법이 나누어짐.



Clustering – Kmeans

Causal-Lab

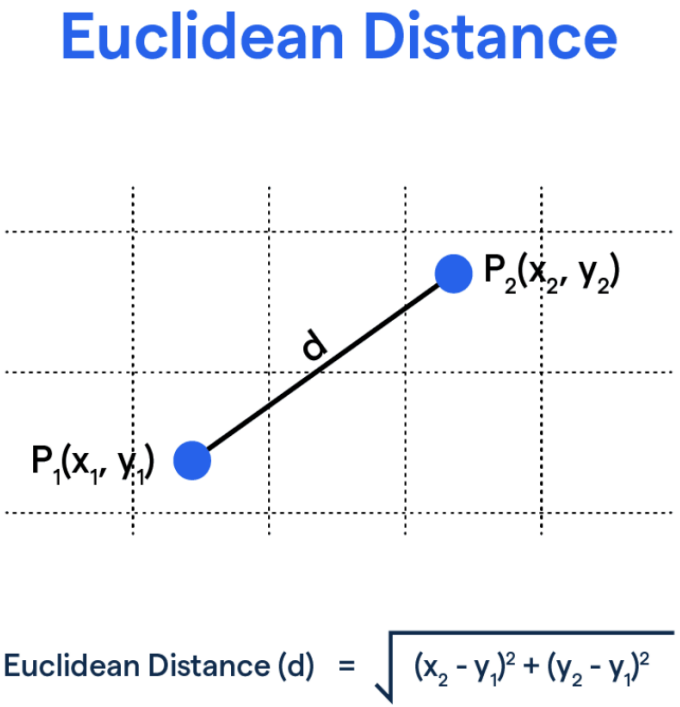
기 정의된 군집 수 기반으로 임의의 centroid와 각 data point간의 거리 계산을 통해 가장 가까운 centroid로 할당하는 방식이며 오직 numerical data만 가능하고 categorical data 경우, 사전에 binary factors로 변환함. 중심점이 더 이상 이동되지 않을 때까지 반복하며, 군집별 centroid와 각 data point간의 분산이 최소화되는 것이 가장 좋은 군집임.



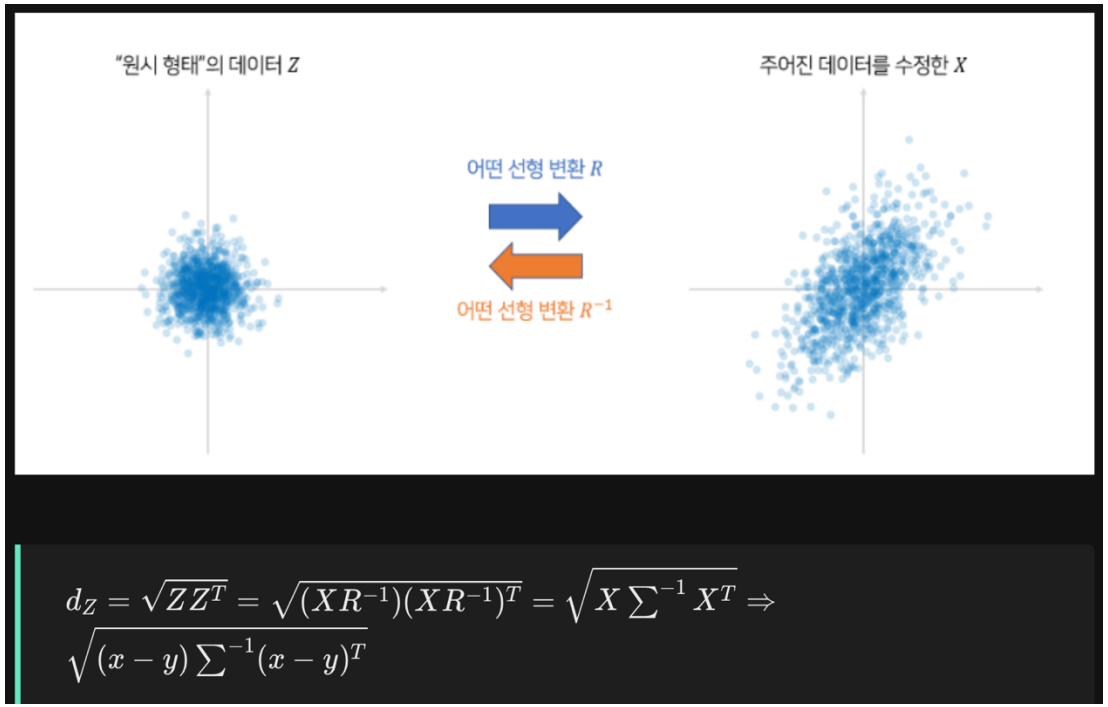
Clustering – Kmeans

기본적으로 euclidean distance를 사용하여 거리를 측정함. 사전에 데이터 정규화가 되어 있지 않으면 Mahalanobis distance 사용함. Mahalanobis distance는 평균과 거리가 표준편차의 몇 배인지 나타내는 것을 의미하며 공분산 행렬이 결국 정규화과정을 의미함.

< Euclidean 거리 공식 >



< Mahalanobis distance 공식 도출 >



Clustering – hierarchical clustering

각 data point가 하나의 군집으로 간주하고 distance가 짧을수록 계속 병합하는 방법이며, 모든 data point를 묶을 때까지 반복 수행함. 사전에 군집수를 지정하지 않아도 되는 장점은 있으나, distance 측정 방식인 linkage는 설정해야 함.

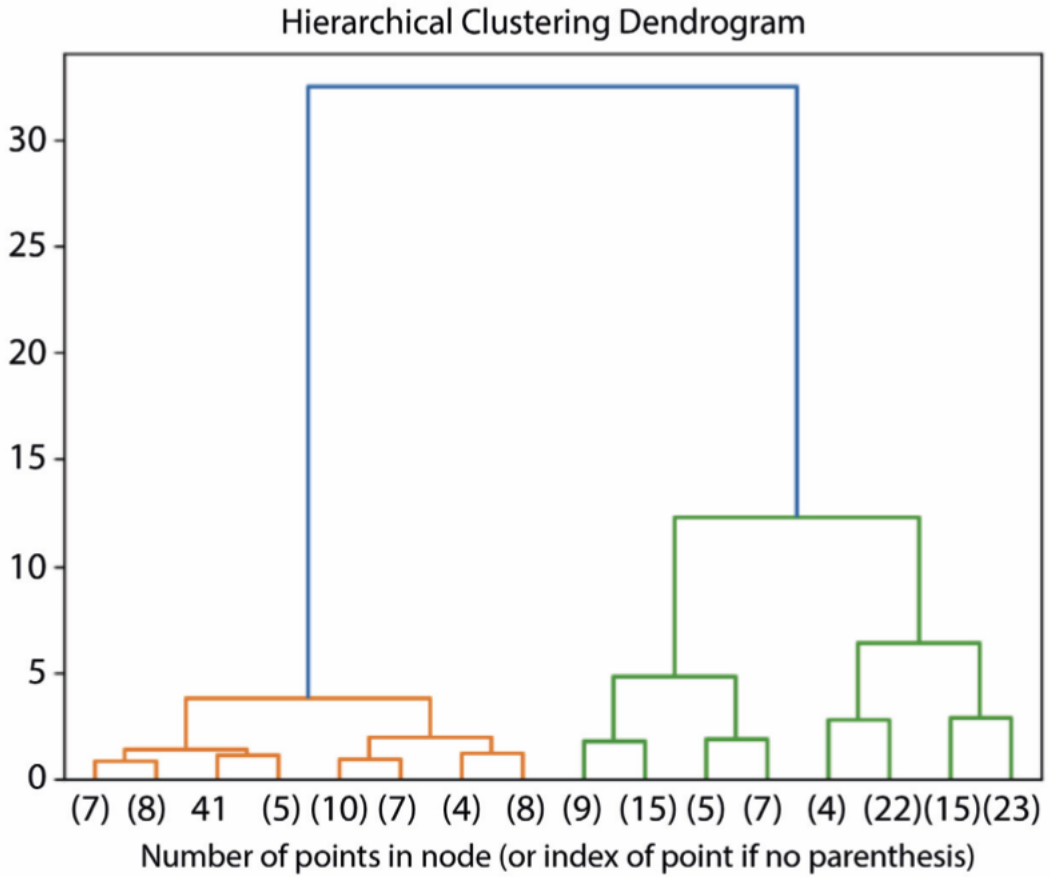
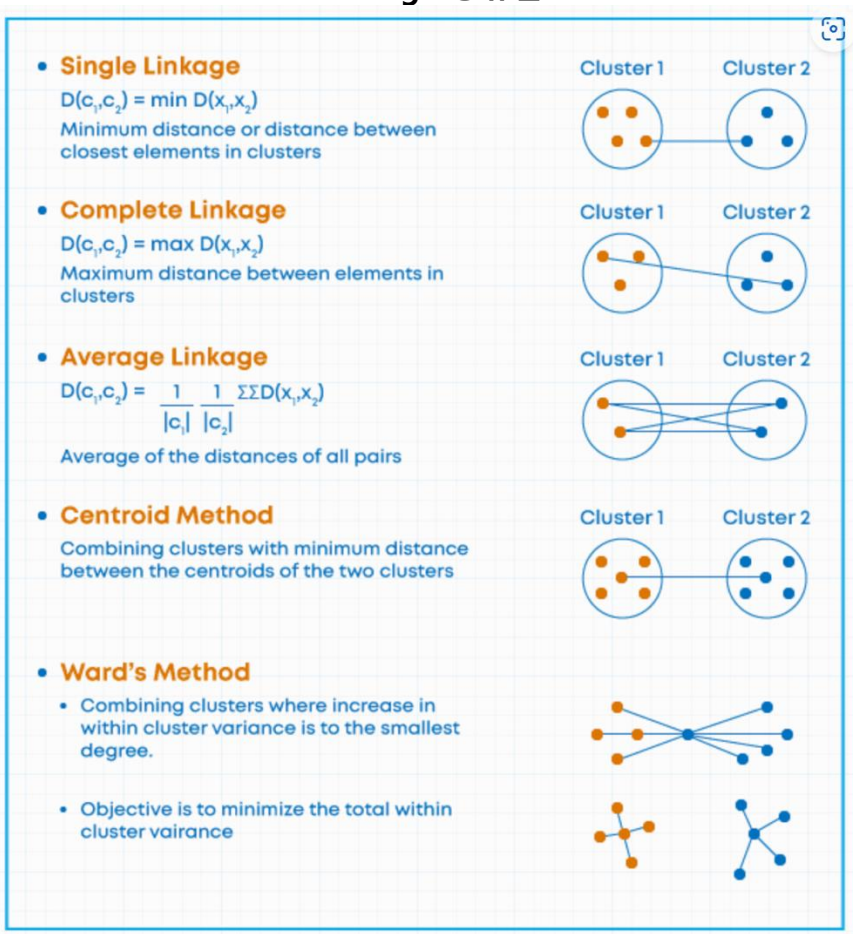


Figure 7.2 – An example dendrogram

< linkage 종류들 >



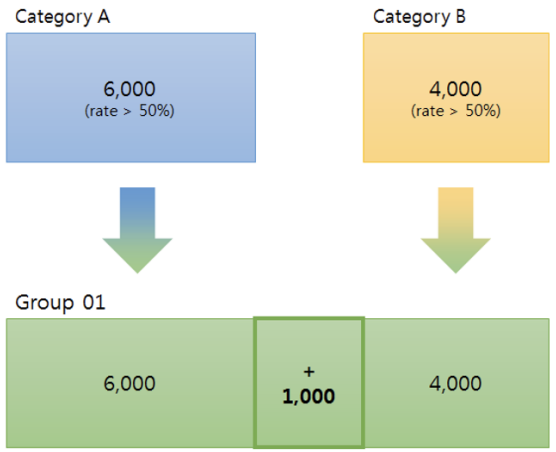
Clustering – hierarchical clustering

Causal-Lab

예시 – 대카테고리기반 고객 군집

Dimension(Feature) Reduction

- 두 카테고리를 합쳤을때 카테고리 영향력이 최대화되는 카테고리를 그룹화
  - 카테고리 영향력 = 특정 카테고리의 클릭 비중이 50%를 초과하는 고객수

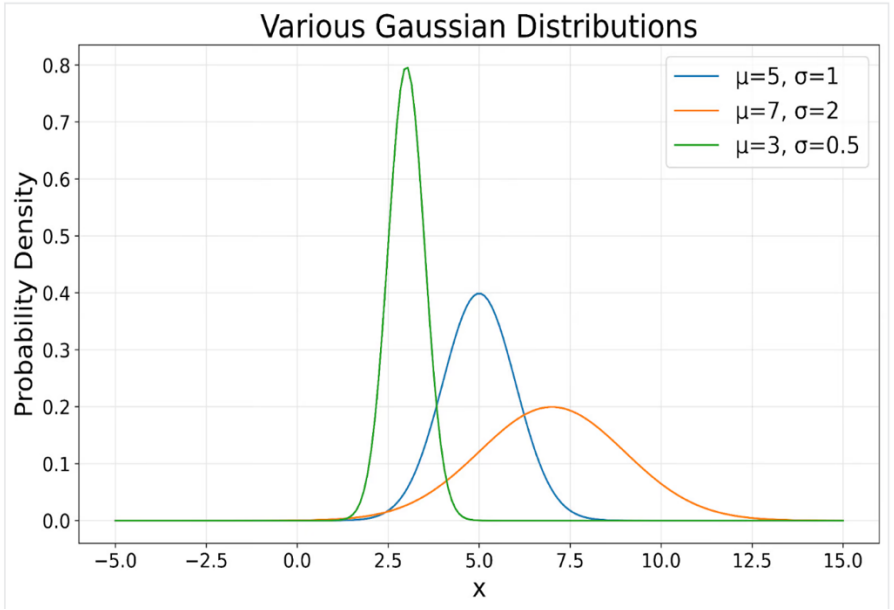


Value = 1000 / (6000 + 4000) = 10%

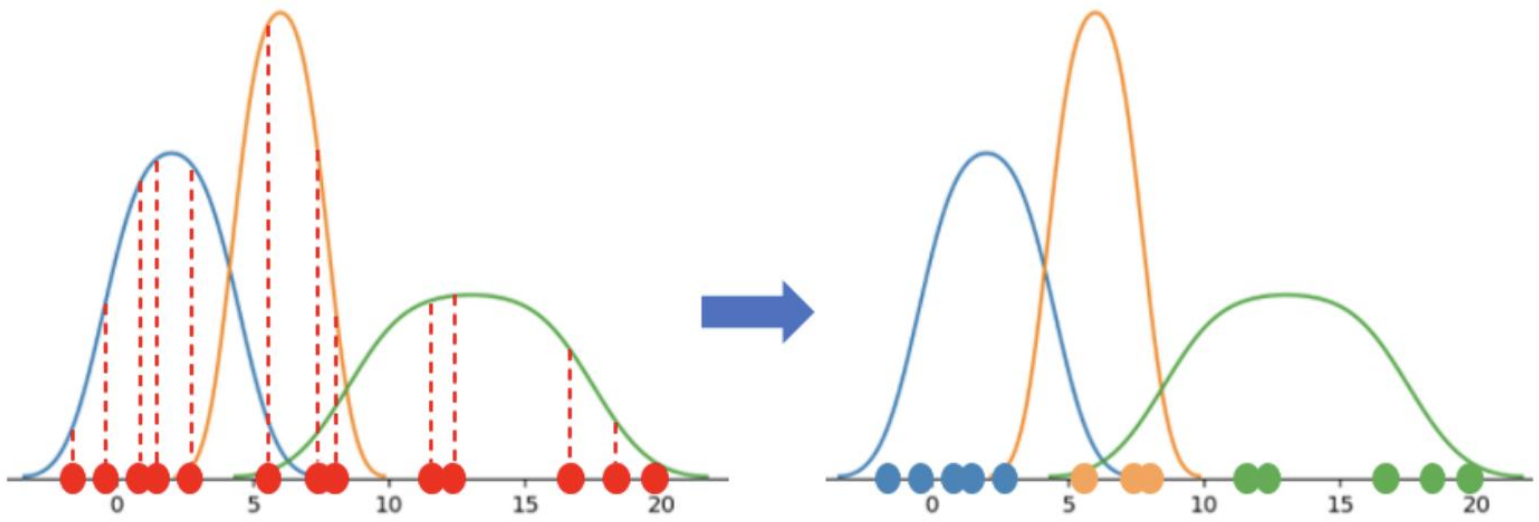
Cluster Mapping Category Info
00 : 모니터/프린터 , 휴대폰 , 노트북/PC , PC주변기기 , 음향기기 , 카메라 , 저장장치
01 : 주방가전 , 주방용품
02 : 자동차용품/블랙박스 , 타이어/오일/부품 , 공구/안전/산업용품 , 캠핑/낚시 , 자전거/보드 , 꽃/이벤트용품
03 : 바디/헤어 , 화장품/향수 , 건강식품/다이어트 , 커피/음료 , 쌀/과일/농수축산물 , 즉석/간식/가공식품 , 세제/구강 , 화장지/물티슈/생리대 , 반려동물용품
04 : 골프클럽/의류/용품 , 남성의류 , 스포츠의류/운동화 , 등산/아웃도어 , 브랜드 캐주얼의류 , 브랜드 남성의류 , 휘트니스/수영 , 구기/라켓
05 : 기저귀/분유/유아식 , 장난감/교육완구/인형 , 유아동의류 , 브랜드 아동패션 , 유아동신발/가방/잡화 , 출산/유아용품/임부복 , 도서/교육/음반 , 디자인/문구/사무용품 , 여행/항공권
06 : 여성의류 , 브랜드 여성의류 , 신발 , 언더웨어
07 : 게임 , 악기/취미/키덜트 , 태블릿
08 : 가방/패션잡화 , 브랜드 신발/가방/잡화
09 : 쥬얼리/시계/선글라스 , 브랜드 시계/쥬얼리 , 수입명품 , 렌탈 서비스
10 : 조명/인테리어 , 가구/DIY , 생활/욕실/수납용품 , 대형가전 , 생활/미용가전 , 계절가전 , 건강/의료용품 , 침구/커튼
11 : 백화점/제화상품권 , e쿠폰/모바일상품권

Clustering – Gaussian mixture model

분석 데이터에 여러 가우시안(정규) 분포가 있다고 가정하고 샘플링을 통해 추출하고,
data point별로 각 분포에 속할 확률을 구한 다음, 가장 높은 확률을 가진 분포에 할당함.
기하학적인 분포에서도 잘 군집이 됨.



< GMM 할당 예시 >



Clustering – 군집 개수

비지도 학습모델의 경우, 군집 개수를 사전에 지정해야 하기 때문에 최적화된 군집 개수를 찾는 것이 매우 중요함. 비지니스 관점에서 임의로 정하는 방법도 있지만, dendrogram을 통해 군집수를 어림 잡거나, elbow기법을 통해 기울기가 완만한 k를 군집수로 간주함.

< 시각화 >

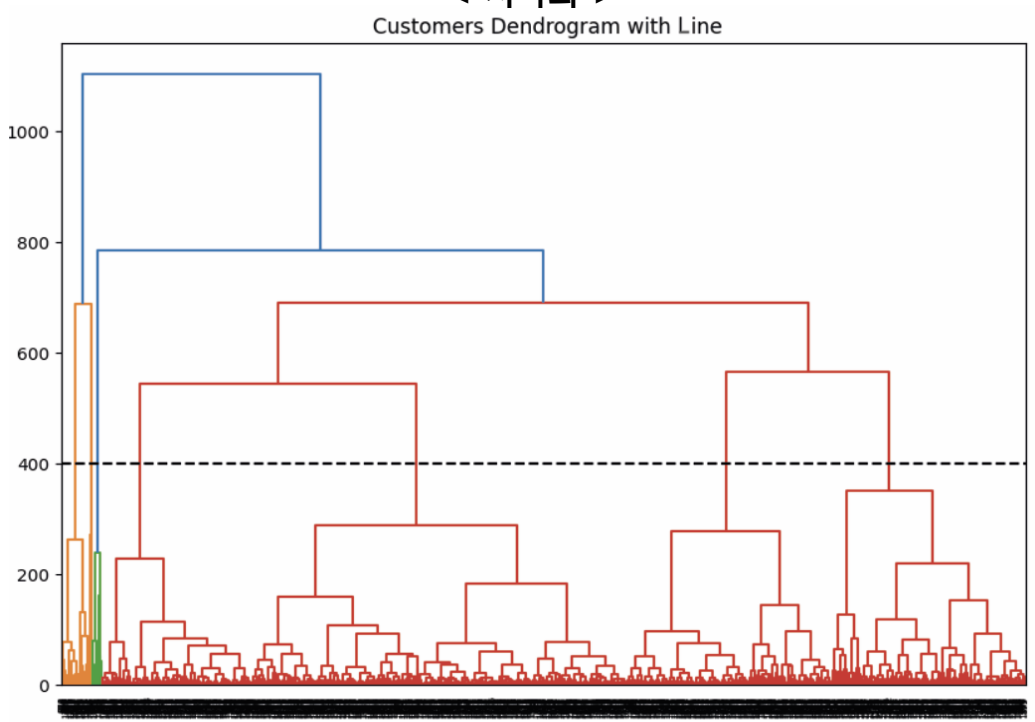


Figure 7.3 – Dendrogram for hierarchical clustering on our dataset

< Elbow 기법 >

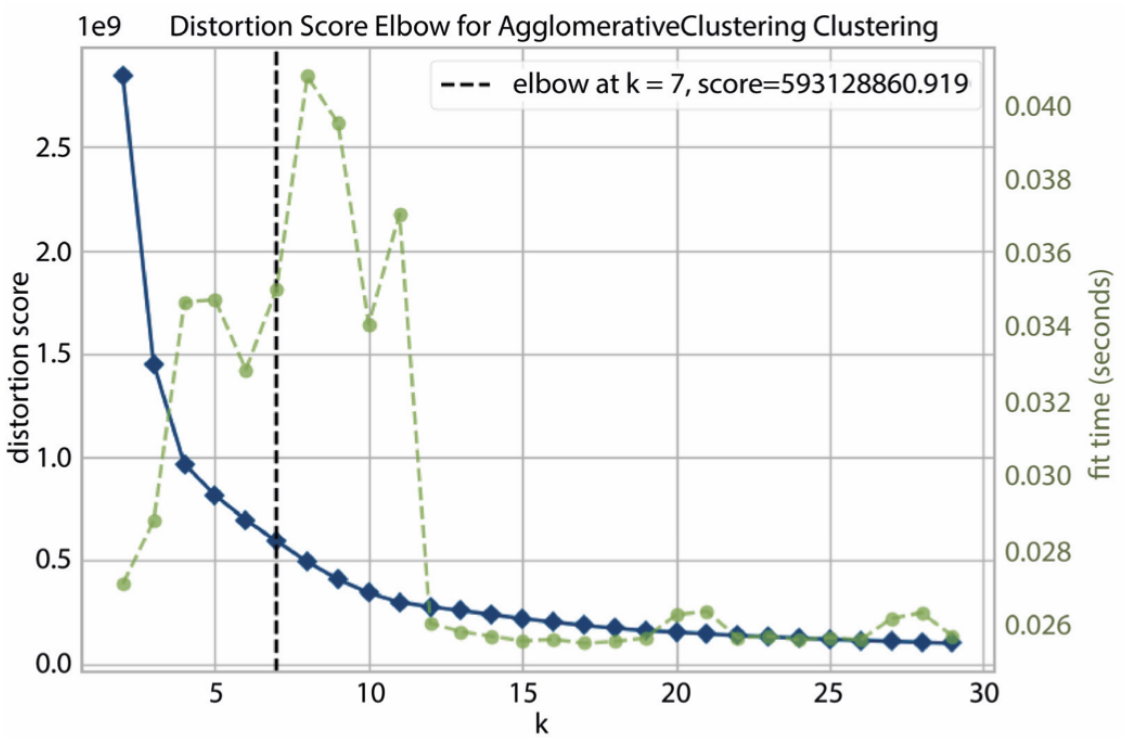
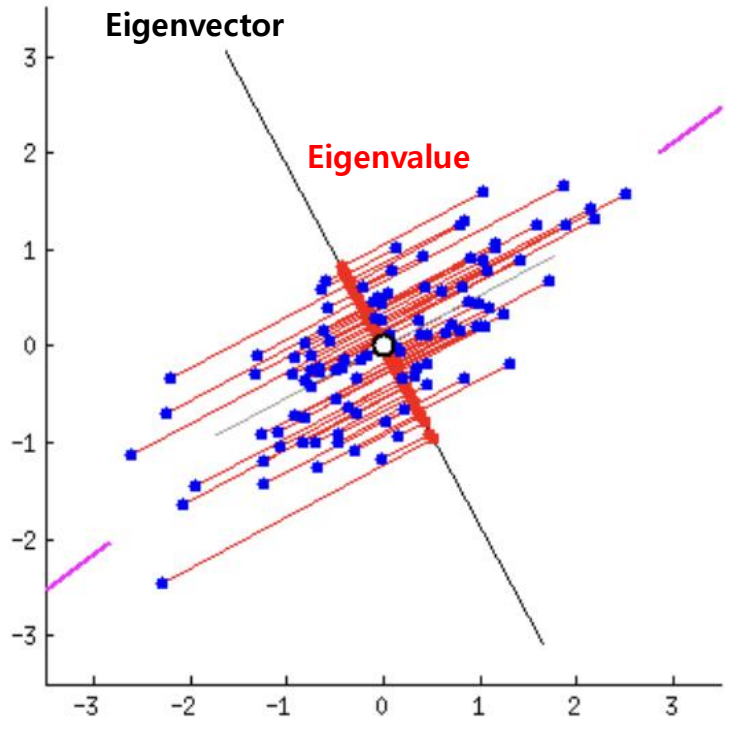


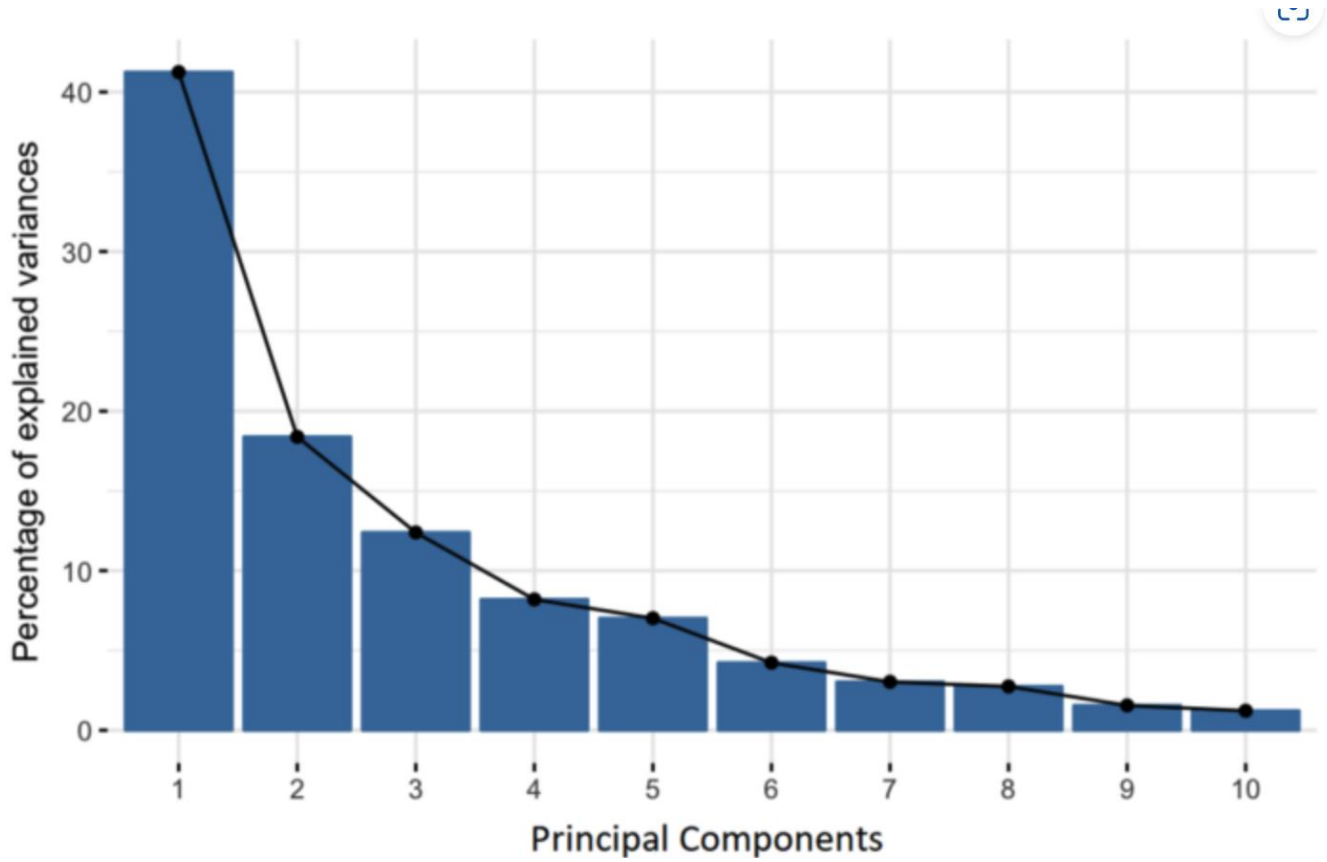
Figure 7.4 – The Yellowbrick KelbowVisualizer

Clustering – 차원 축소(PCA)

방대한 데이터와 여러 속성기반으로 기법들을 수행할 경우, 엄청난 시간/리소스 비용이 발생함.
차원 축소기법을 활용하여 저차원으로 변환하고 이를 기반으로 elbow기법들을 다시 수행함.
PCA 경우, variance가 높은 순으로 기준이 되는 설명력까지 속한 principal components로 차원을 줄임

< PCA >

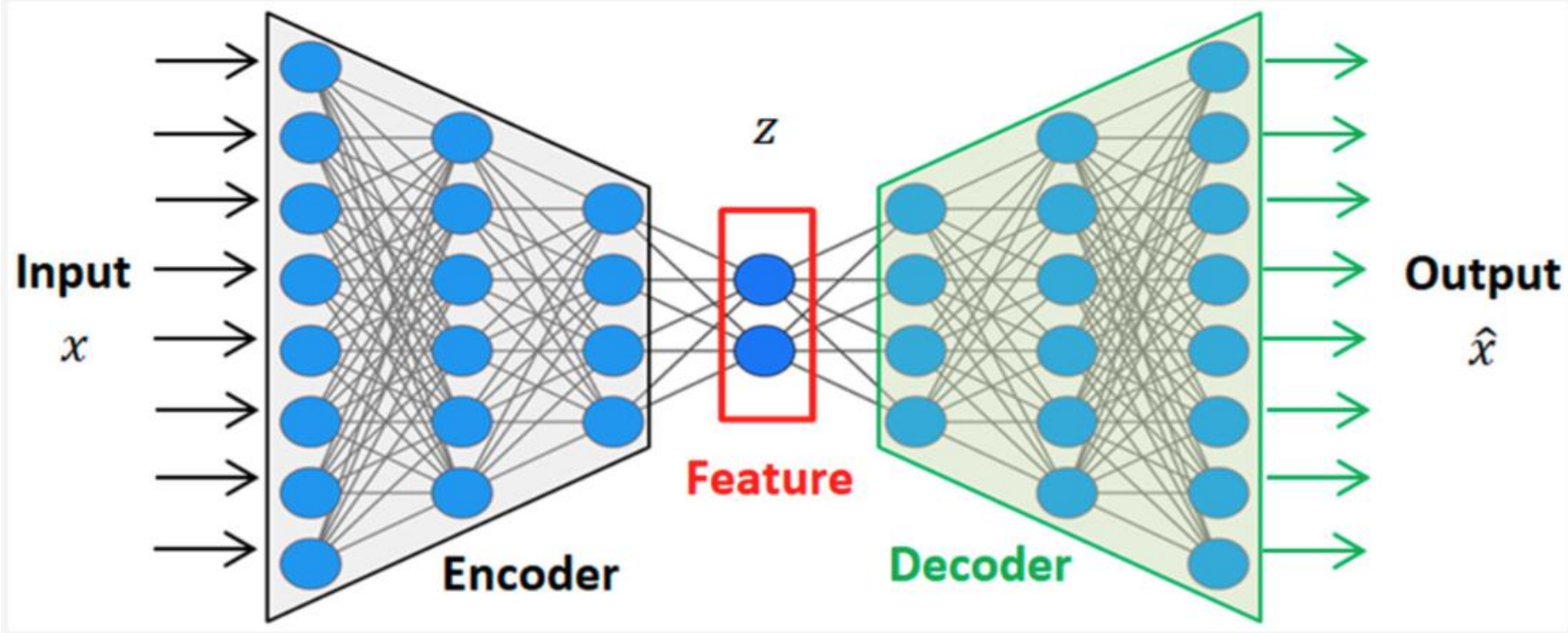




Clustering – 차원 축소(Autoencoder)

Causal-Lab

Autoencoder 경우, 학습 데이터 차원에서 점차 줄여가며 압축하는 encoder 부분과 다시 차원을 늘려 복원하는 decoder 부분이 있음. 이때 가장 저차원으로 학습된 feature를 추출하게 되면 고차원에서 저차원으로 줄여준 것과 동일한 효과이며, 이를 기반으로 elbow나 시각화를 통해 군집개수를 지정함



Classification – Random Forest

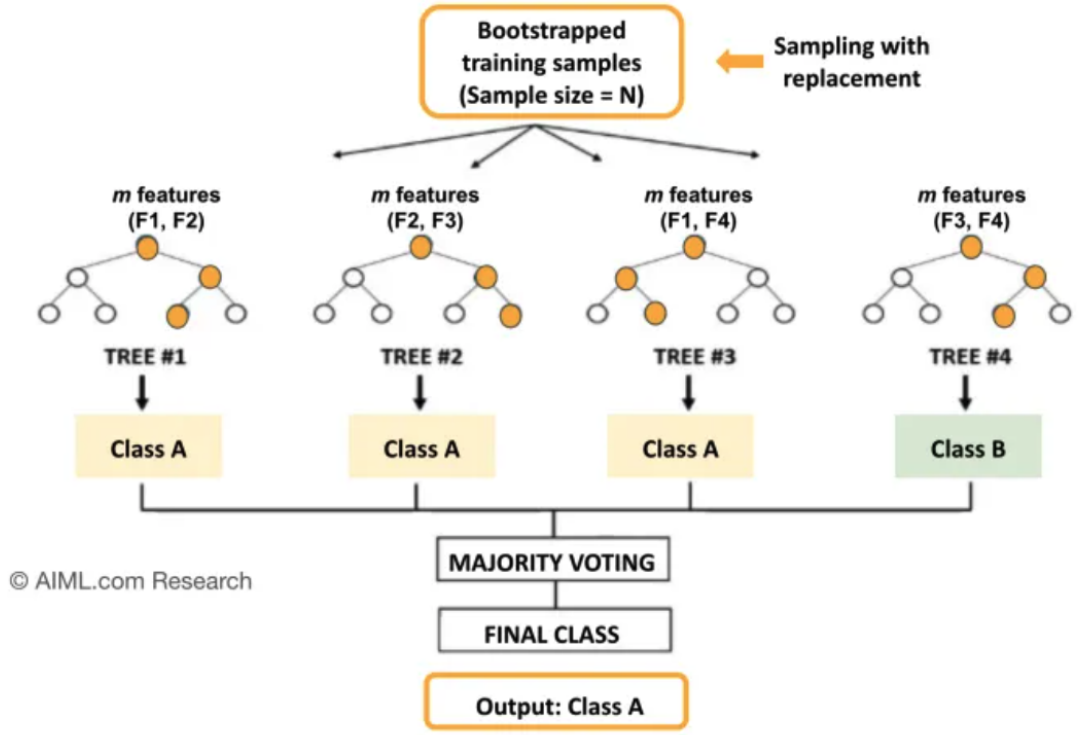
Causal-Lab

Decision Tree 모델의 Overfitting 단점을 해결한 모델이며, 학습 데이터의 subset 기반으로 작게 트리를 여러개 구성하고 이들을 하나로 합쳐 label를 분류함. Outlier or non-linear 데이터에서도 결과가 좋음

< Random Forest >

Training Data
(Sample size, N= 6, No. of features, F = 4)

F1	F2	F3	F4	Y
2.1	0	400	-9	A
3.0	1	890	-42	B
2.2	1	929	0	B
4.0	0	324	-23	A
3.5	1	333	-15	A
6.0	0	215	-9	A



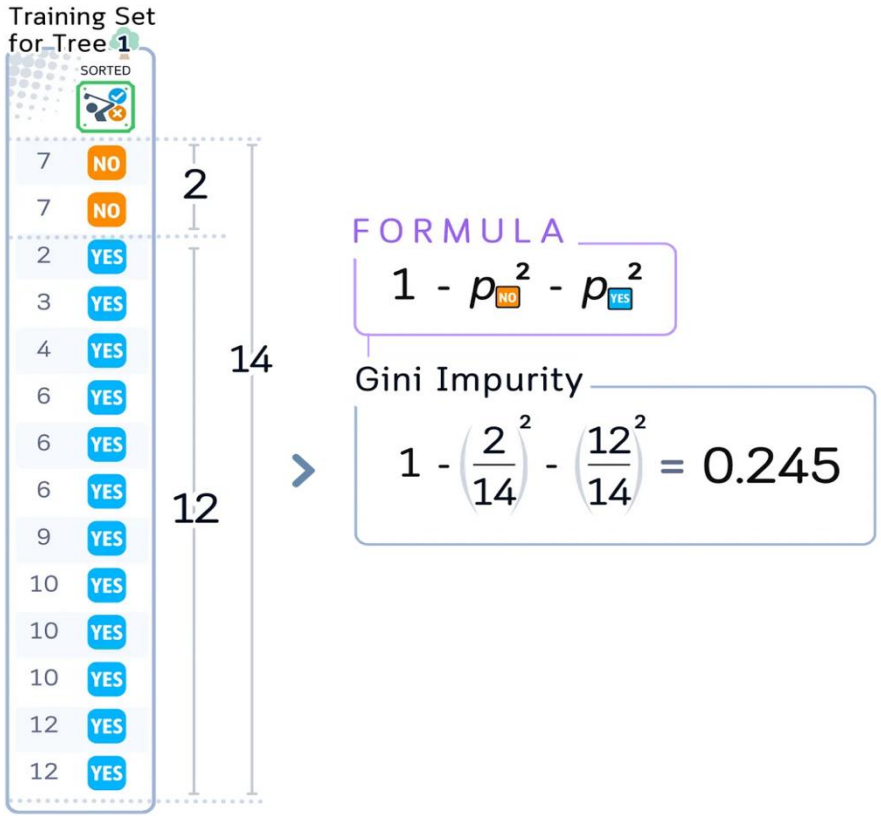
Key parameters of Random Forest Model are: (a) Number of trees , (b) Maximum depth of the trees (c) Size of the random subset of features
In this example, No. of trees = 4, Depth = 2, and Feature subset size, m = 2 (no. of features/2)

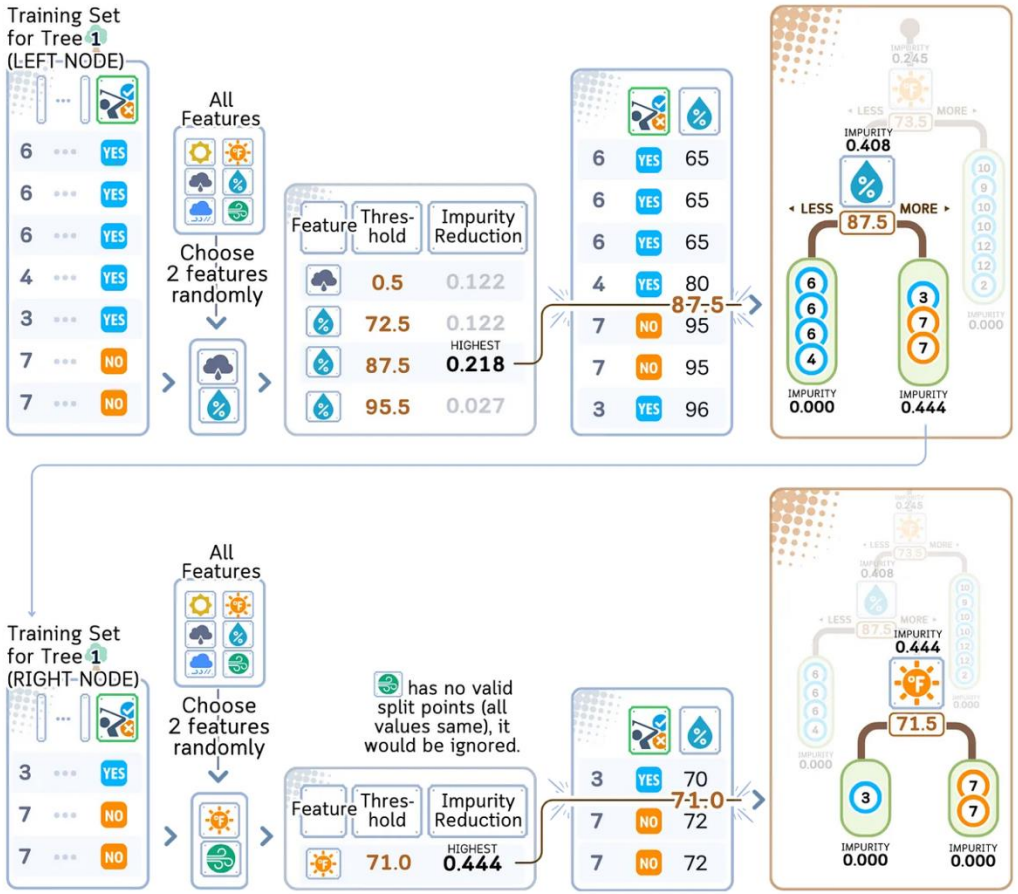
Classification – Random Forest

Causal-Lab

Gini Impurity기반으로 가장 감소가 큰 포인트를 기반으로 데이터셋을 나누는 방식임

< Random Forest >





Classification – Confusion Matrix

Causal-Lab

Classification 모델의 성능평가로 얼마나 잘 분류되는지 여러 지표로 검증함.

		Predicted condition	
	Total 8 + 4 = 12	**Cancer** 7	**Non-cancer** 5
Actual condition	**Cancer** 8	6	2
	Non-cancer 4	1	3

Figure 7.9 – An example confusion matrix

- **Accuracy**: The proportion of true results (both true positives and true negatives) in the total number of cases examined:

$$(TP + TN) / Total = (6 + 3) / 12$$

- **Precision**: The proportion of true positive results in the number of cases that were predicted as positive:

$$TP / (TP + FP) = 6 / (6 + 1)$$

- **Recall (sensitivity)**: The proportion of true positive results in the number of cases that are actually positive:

$$TP / (TP + FN) = 6 / (6 + 2)$$

- **Specificity**: The proportion of true negative results in the number of cases that are actually negative:

$$TN / (TN + FP) = 3 / (3 + 1)$$

RFM

Causal-Lab

Recency, Frequency, Monetary 3개의 지표기반으로
Independent 와 Sequential Sorting 방식으로 고객을 세분화함

Independent Sorting

- 각각 equal-sized groups
- 125 groups (5 Recency x 5 Frequency x 5 Monetary)
- 작은 데이터 셋이거나 imbalanced distribution 경우, 제대로 나눠지지 않음

Sequential Sorting

- Nested 방식
- Monetary 기반으로 그룹을 나눈 다음, Recency, Frequency순으로 다시 세분화함
- 작은 데이터 셋 유리함

RFM - 타겟팅

Causal-Lab

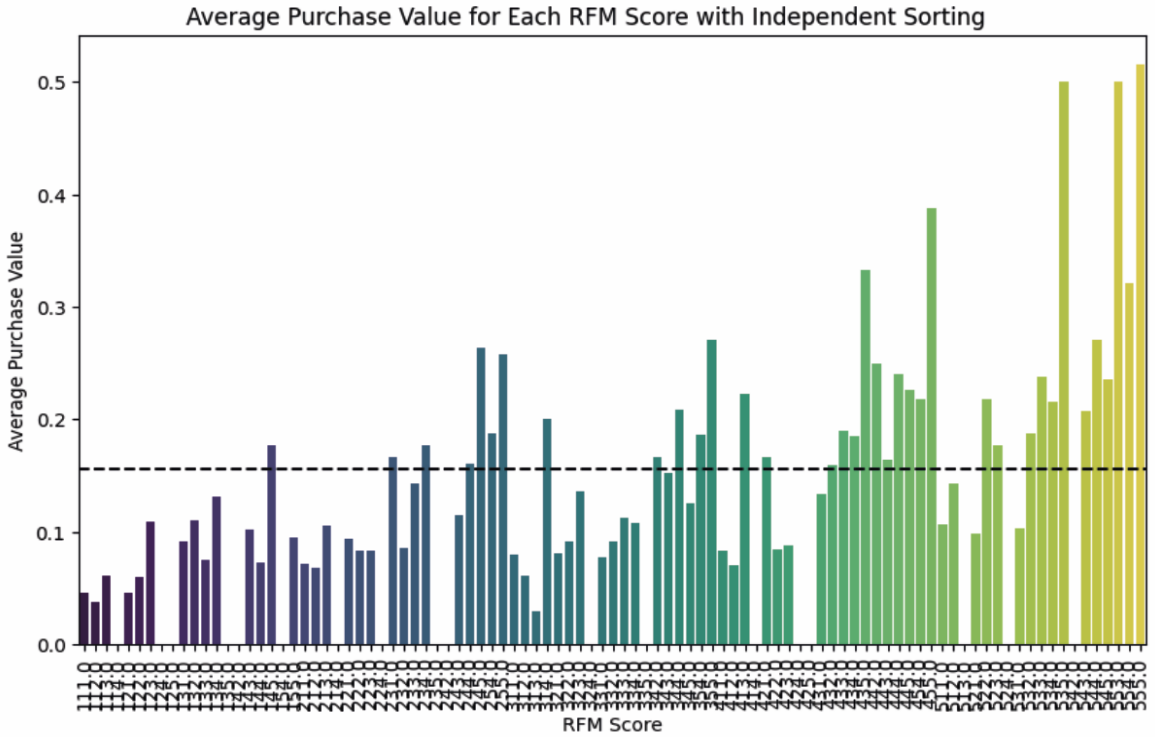
Return on marketing investment or ROMI 기반으로 타겟팅함

$$Breakeven = \frac{Cost\ to\ do\ campaign}{Profit\ from\ single\ sale}$$

- Average purchase amounts from the marketing campaign: 40 USD
- Average cost of the goods: 19.20 USD (equivalent to a 52% gross margin)
- Average cost of shipping the product: 6 USD
- Average cost of the marketing campaign: 2 USD

So this means that the average gross profit is simply 40 – 19.20 = 20.80 USD. Discounting the cost of shipping and the cost of the marketing campaign, we get 20.80 – 6 – 2 = 12.80 USD of profit per customer that made a purchase.

Now, the question becomes, what percentage of customers in each RFM cell needs to purchase if we send the catalog to all customers, which is simply 2/12.80, giving us a breakeven rate of 15.625%?



RFM – 타겟팅

Causal-Lab

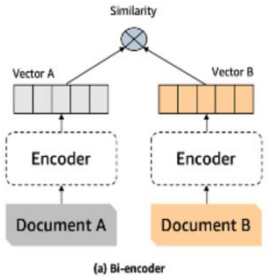
예시 – 브랜드 기반 증강 고객 세분화 및 타겟팅

상품-키워드 pair

- 상품 - 소카테고리
- 상품 - 쿠폰명
- 상품 - 검색어
- 상품 - 브랜드
- 상품 - 제조사

중요도만큼 pair 증강

Bi_encoder기반 모델 생성 및 상품별 Top5키워드 추론



상품 번호	상품명	추론 키워드	유사 점수
3406275901	백설 콩식용유 1.8l 5개입	콩기름/대두유, 콩기름, 백설, 식용유...	0.96,0.80,0.72,0.54...
2994206536	CJ 미트볼로냐스파게티2인(625g) x 3개	CJ제일제당,스파게티/파스타,백설,스팸..	0.94,0.20,0.008,0.001..
3072435705	햇반 흑미밥 210g 3개입x8개(총24개)	CJ제일제당,즉석조리/볶음요리,백설,햇반즉석밥..	0.85,0.04,0.03,0.01...
3576483390	비비고 백서 너비아니 560g 3개	백설,떡갈비,CJ제일제당,비비고...	0.66,0.61,0.37,0.009...
...	...	...	...

상품(Document A) - 키워드(Document B) 각각 임베딩한 후 유사도를 계산

연관 키워드 추출

CJ제일제당(제조사) > 백설(하위브랜드)

'백설' 동선라인 키워드
(cutoff : 0.8이상)

콩기름/대두유,
콩기름, CJ제일제당 ...

점수가 높을수록 해당 상품과 연관성 높음

타겟키워드에 속한 상품 대상 고객 세분화

세그먼트명	대상 고객수	평균 주문건수	평균 주문수량	평균 주문액
S	12,009	23.5	44.9	461,271
A	165,860	4.3	6.9	119,038
B	165,962	1.9	2.7	47,051
C	166,038	1.3	1.6	27,575
D	166,161	1.2	1.5	13,924
E	346,220	N/A	N/A	N/A

잠재 고객

가짜연구소

Causal-Lab

감사합니다.