**차원축소(PCA)**

**Theory**

분석에 사용할 데이터의 변수가 너무 많을 경우 PCA 등을 사용하여 변수를 축소하고, 데이터를 가장 잘 설명할 수 있는 적절한 변수들만 선택하여 분석에 사용함. 결과적으로 데이터의 특징을 가장 잘 나타내는 주요 특징만 추출하는 절차(Feature Extraction)

PCA는 여러 변수 간의 상관관계를 이용하여 이를 대표하는 주성분을 추출하여 차원을 축소하는 기법. PCA는 제일 먼저 가장 큰 데이터 변동성(Variance)을 표현하는 첫 번쨰 벡터 축을 생성하고, 두 번째 축은 이에 직각이 되거나, 그 다음으로 가장 많이 설명할 수 있는 벡터를 축으로 생성함. 이렇게 생성된 벡터 축에 원본 데이터를 투영하여 차원을 축소. 각 주성분은 서로 독립인 것(cor=0)이 원칙.

PCA하기 전에 표준화를 해야함. 이렇게 해야 변수 간의 단위 차이로 인해 벌어지는 문제점을 방지할 수 있음.

**장점**

차원축소를 사용하면 다차원 특성으로 인해 발생하는 희소구조(Sparser)를 저차원으로 축소시켜 학습모델의 성능을 높일 수 있으며, 시각적 표현, 메모리 효율 등에 대해서도 이점이 크다. 특히 선형회귀와 같은 선형 모델의 경우 입력 변수간의 다중공선성 문제를 해결하여 모델 성능의 큰 개선을 이룰 수 있다.

**핵심**

주성분 기여율(Proportion of Variance) : 주성분별로 원본 데이터의 변동성을 얼마나 반영하고 있는지로 (주성분 변수의 분산) / (원 변수의 총 분산)으로 계산함.

첫 주성분부터 차례대로 주성분기여율을 합쳐 누적 기여율(Cumulative Proportion)이 85% 이상이 되면 해당 지점까지를 주성분의 수로 결정함.

스크린 산점도(Scree Plot)을 통해 고유치가 급격히 완만해지는 지점의 바로 전 단계로 주성분의 수를 선택.

**전략**

상관관계 분석 -> Befor EDA -> Standardization -> 차원 수 튜닝(스크린 산점도, 주성분 기여율, 누적 기여율) -> PCA -> After EDA

다변량 중 2개를 골라서 Before EDA 시각화를 하고, 데이터셋에 표준화 및 PCA를 수행. 그리고 주성분 기여율, 누적 기여율을 구하고 스크린 산점도를 그림. 마지막으로 After EDA 시각화를 그린 후, 본 차원축소 작업이 전체 데이터셋을 가장 잘 표현하는 주요 특성을 추출하여 추후 이루어질 모델 학습에 기여할 것으로 기대할 수 있다. 로 마무리.