Machine Learning – Preprocessing & Dimensionality Reduction

Index

preprocessing

전처리

누락된 데이터

- 수집과정에 오류가 있거나 측정 방법 적용이 불가능할 때 발생한 사용할 수 없는 값
- NaN (not a number)
- NULL (빈 값)

값을 대체하는 방법 (보간법)

- 평균값 대체
- 최빈값 대체
- 중앙값 대체

범주형 데이터

- 순서, 크기를 나타낼 수 있는 데이터 (XL > L > M)
- 순서가 없는 데이터 (빨강, 파랑, 초록)

데이터 인코딩 방법들

- 특성 간의 산술적 차이를 활용한 값 매핑
- 레이블의 개수만큼 0부터 1씩 부여하는 레이블 인코딩
- 클래스 수만큼 차원수를 늘리는 원핫 인코딩

스케일링

- 특성들간의 스케일을 맞춰주는 작업

스케일링 방법들

- 최소값과 최대값을 통해 0과 1의 범위로 조정하는 MinMaxScaler
- 데이터를 표준화하는 StandardScaler
- 이상치가 많이 포함된 작은 데이터셋을 다룰때 유용한 RobustScaler

$$X_{\text{scale}} = \frac{x_i - x_{\text{med}}}{x_{75} - x_{25}}$$

- robust scaler

Feature

특성

특성 선택 (Feature Selection)

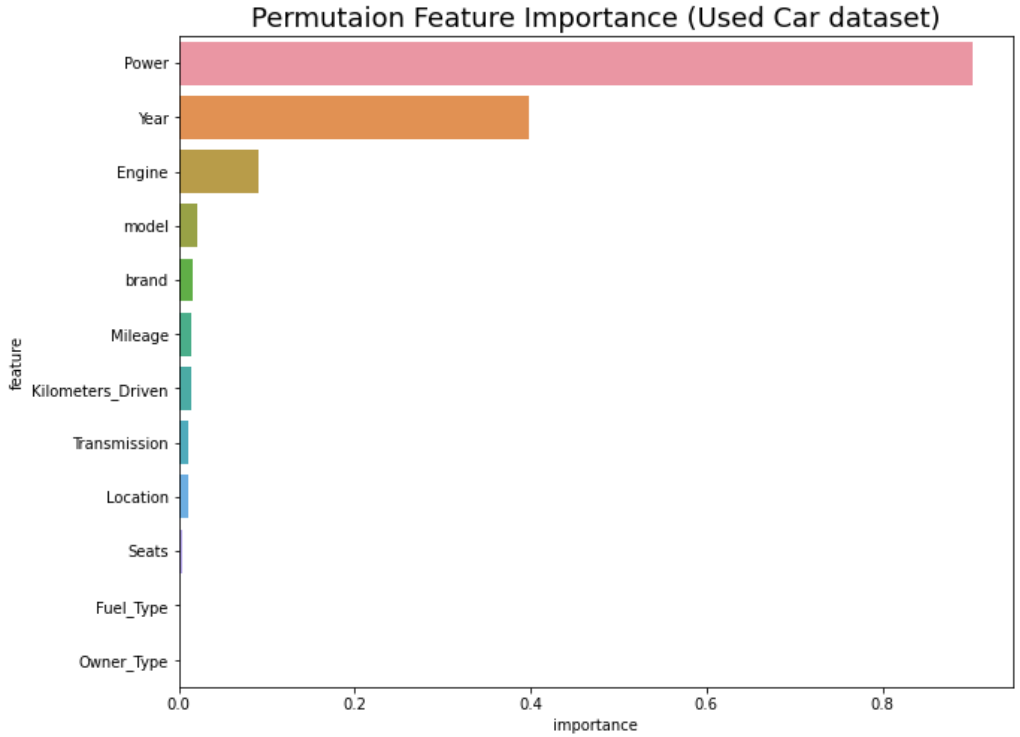
- 모델 학습 후 테스트 데이터셋보다 학습 데이터셋에서의 성능 차이가 크다면 overfitting에 대한 신호
- 대체로 훈련 데이터셋에 비해 모델이 복잡할 경우 발생

Overfitting 해결 방법

- 학습 데이터를 추가
- 규제를 통한 모델의 복잡도 제한
- 파라미터 개수가 적은 모델 선택
- 데이터의 차원 축소

Feature Importance (in RF)

- 랜덤 포레스트와 같이 앙상블에 사용한 모든 Decision Tree에서 계산한 평균적인 불순도 감소로 특성 중요도를 측정 가능



피처 중요도를 활용한 피처 선택

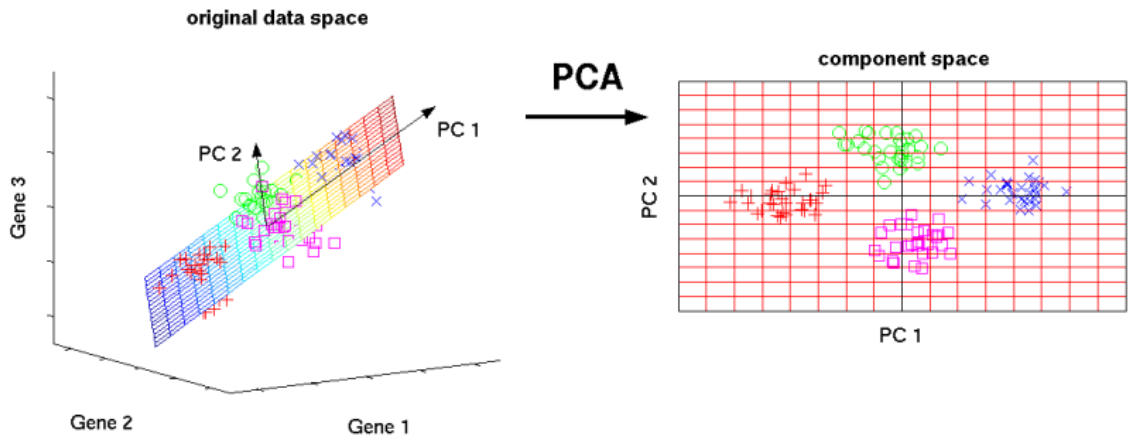
- 사용자가 설정한 threshold를 활용해 Feature Importance의 값이 해당 임계값 이상일 때 피처 선택

Dimensionality Reduction

차원 축소

주성분 분석

- Principal Component Analysis, PCA
- 관련 있는 정보를 유지하면서 데이터를 압축하는 방법
- 줄어든 차원 수로 계산 효율성을 향상시키고, 차원의 저주 문제를 감소시켜 모델 성능 향상을 기대할 수 있음



주성분 분석 계산 순서

- d차원 데이터셋을 표준화 전처리 (Standard Scaler)
- 공분산 행렬을 생성하여 고유벡터와 고유값으로 분해
- 고윳값이 가장 큰 k개 고유벡터를 선택한다 (k는 d보다 작은 수)
- k개의 고유벡터로 만들어진 projection matrix W를 만들고 원본 데이터셋 X과 행렬곱하여 새로운 X`을 생성

커널 PCA

– 데이터가 선형이 아니라 비선형일 때도 차원 축소를 할 수 있을까?
– 커널 PCA의 경우 기존 데이터를 고차원 공간으로 매핑 후 선형 분류기로 분류할 수 있는 저차원 공간으로 데이터를 투영한다.