2주차 예비 보고서

전공: 컴퓨터공학 학년: 2학년 학번: 20221559 이름: 박준우

1. Feature Engineering이란 머신 러닝 알고리즘이 작동할 수 있도록 하는 Feature들을 만드는 과정이다. 이것을 어떻게 하느냐에 따라 같은 알고리즘이어도 성능이 크게 바뀔 수 있기 때문에, 이용하는 입장에서 성능을 크게 개선시킬 수 있는 중요한 과정이다. 어떻게 처리할 것인가는 데이에 대한 이해를 필요로 하기 때문에, 처리할 수 있는 수단들의 이해를 바탕으로, 사용할 수 있는 방식 중 어떤 것을 이용해야 할 지 잘 판단하는 것이 중요하다.

1) 아웃라이어 (이상치) 처리

실제로 주어진 데이터에는 전체의 특성을 파악하는 데 있어 방해가 될 수 있는 요소가 포함되어 있을 수 있다. 요인으로는 데이터 수집 과정에서 오류가 발생하거나, 데이터 자체에서 포함하고 있는 경우 등이 있다.

아웃라이어(이상치)는 일반적인 추세를 따르지 않고 멀리 떨어져 있는 값을 말한다. 물론 아웃라이어에 잘 대응하는 모델도 있지만, 많은 알고리즘들은 아웃라이어로 인해 모델이 일반적인 값들을 잘 대표하지 못하는 추세를 찾을 수 있다. 따라서 이런 아웃라이어를 제거해 주는 것으로 모델의 성능이 개선될 수 있다. 이런 값들은 상대적이다. 즉, 어떤 데이터를 아웃라이어로 분류할 것인가에 따라 아웃라이어로 판단되는 값이 달라질 수 있다. 분류 기준으로 사용할 수 있는 방법은 여러가지가 있는데, 통계적 분포에서 주변값들을 제거할 수도 있고, 직접 그래프를 그려 추세에서 벗어난 데이터를 파악한 후, 기준을 직접 설정할 수도 있다.

2) Feature Encoding

Feature Encoding은 범주형 데이터를 모델이 처리할 수 있도록 수치형 데이터로 바꾸는 과정이다. 예를 들어, 성별이 male, female로 되어 있으면 이것을 0-1로 바꿔주는 것 등이 있다.

할 수 있는 방법도 여러가지가 있다. 각 데이터들에 1대1 대응으로, 해당 요소가 있으면 1, 없으면 0을 가지는 새로운 행을 만드는 One-Hot Encoding, 또 여러 요소들을 각각 고유한 정수에 매칭시켜 나타내는 Label Encoding 등이 있다.

3) Feature Generation

주어진 Feature가 모델이 판단하기 어려운 형태로 되어 있을 때, 여러 요소들을 조합하거나, 변환해서 더 직관적으로 파악하기 쉬운 형태로 바꿔주는 것이다.

예를 들어 키와 몸무게 데이터가 있을 때, 이를 조합해 체질량지수로 변환해 주는 것이나, 1주차에 진행한 집값 데이터의 위도, 경도 정보를 이용해 지역구로 변환해 줄 수 있겠다.

2. 결측치(누락된 값)가 포함되어 있으면 생길 수 있는 문제는 여러가지가 있다. 모델에 따라선 NULL이나 NaN값 등에 대해서는 작동이 안 되거나, 실제와 다른 방향으로 해석할 수 있는 가능성이 있기 때문에, 이를 처리해 줌으로 모델의 정상적 작동을 보장해 줄 수 있다.

처리 방식도 결측값이 얼마나, 어떤 데이터에 존재하는가에 따라 여러 방법이 있다. 결측값의 수가 적은 경우,, 결측값이 포함된 데이터 행 전체를 삭제할 수 있다. 이런 결측값들을 다른 데이터들의 평균, 중앙값, 최빈값 등의 대표값으로 채울 수도 있고, 결측값이 나왔다는 별도의 특징으로 처리할 수도 있다.

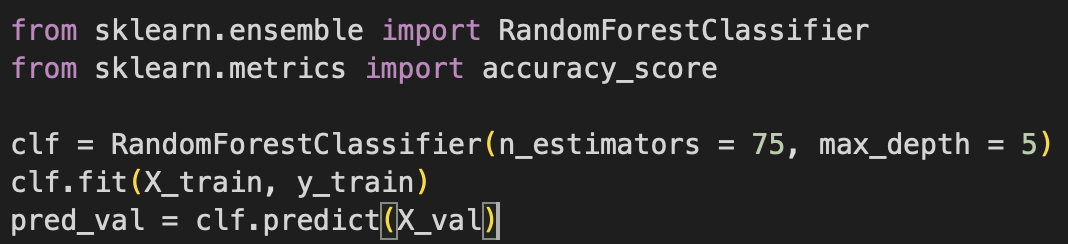
3. Random Forest – classification / regression

Random Forest는 앙상블 학습의 일종으로, 여러 개의 Decision Tree를 생성하고, 이를 결합하여 최종 예측을 수행하는 강력한 머신러닝 알고리즘이다. 이 모델은 데이터의 일부를 샘플링하여 여러 개의 트리를 생성하는 배깅(Bagging) 기법을 사용하며, 각 트리의 예측을 회귀 모델에서는 평균으로, 분류 모델에서는 다수결 투표를 통해 최종 결과를 도출한다. Random Forest는 높은 예측 정확도를 제공하며, 과적합(overfitting)을 방지하는 데 효과적이다.

특징적으로, Random Forest는 랜덤성을 활용하여 트리의 다양성을 높인다. 이를 위해 각 트리 생성 시, 데이터 샘플을 랜덤하게 선택하고, 각 분기에서 사용하는 변수(feature)도 랜덤하게 선택한다. 이러한 방식은 모델의 편향을 줄이고 분산을 낮추어 일반화 성능을 향상시키는 데 기여한다. 또한, Random Forest는 변수의 중요도를 평가할 수 있는 기능을 제공하며, 이를 통해 데이터 분석 및 변수 선택에 유용하게 활용할 수 있다.

먼저, 학습에 사용할 입력 데이터와 레이블을 준비하고, RandomForestClassifier 또는RandomForestRegressor를 호출하여 모델 객체를 생성한다. 이때 n\_estimators (트리 개수),

max\_depth (트리 최대 깊이)와 같은 하이퍼 파라미터를 설정할 수 있다. 준비된 데이터를 fit() 메서드를 통해 모델에 학습시키고, 학습된 모델로 새로운 데이터에 대해 predict() 메서드를 사용하여 예측을 수행한다.



예시 코드