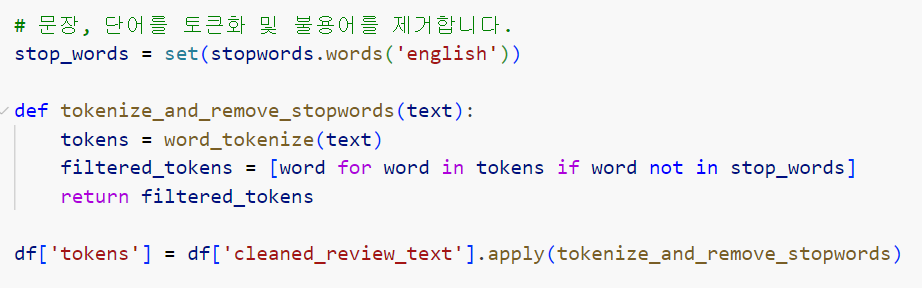
**<기말고사 대체 과제>**

201920835 소프트웨어전공 김기범





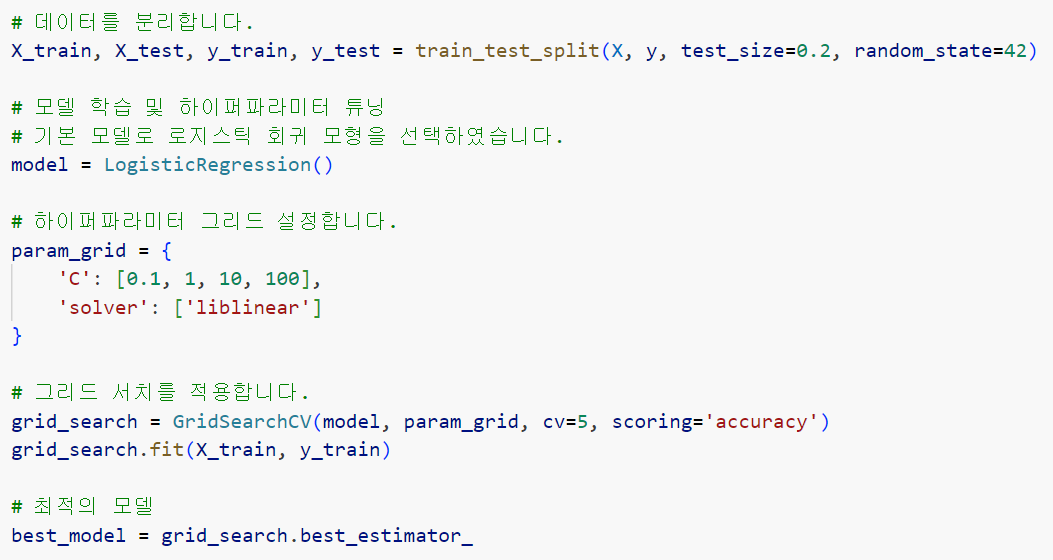
자연어 처리를 위해 nltk 데이터를 다운로드하고 아마존 신발 제품 리뷰 데이터를 로드하였습니다. 그리고 리뷰 텍스트의 결측값을 빈 문자열로 대체하였습니다. HTML 태그, URL, 특수 문자 등을 제거하고 모든 텍스터 데이터를 소문자로 변환 후 텍스트 데이터를 정리하였습니다.

텍스트 데이터를 단어 단위로 분할하고, 의미가 없는 불용어를 제거하였습니다.

단어의 다양한 형태를 표준 형태로 통합하기 위해 스테밍과 표제어 추출을 수행하였습니다.

리뷰 텍스트 데이터를 TF-IDF 벡터화하여 머신러닝 모델이 이해할 수 있는 수치형 데이터로 변환하고, 각 리뷰의 감성을 나타내는 레이블을 설정했습니다.

레이블은 설정은 감성 분석의 목표인 긍정 리뷰와 부정 리뷰를 구분하기 위해 평점이 4이상인 경우 긍정 리뷰, 그렇지 않은 경우 부정 리뷰로 간주하였습니다.

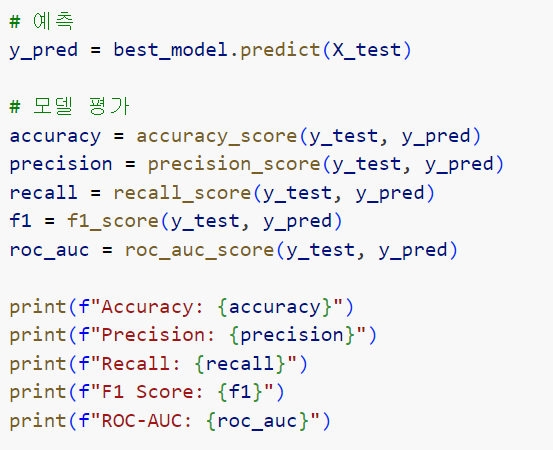
데이터를 학습 세트와 테스트 세트로 분할하고, 로지스틱 회귀 모델을 학습시키며, 최적의 하이퍼파라미터를 찾기 위해 그리드 서치를 수행하였습니다.

전체 데이터의 20%를 테스트 세트로 사용하였고, 재현 가능한 결과를 얻기 위해 랜덤 시드를 설정하였습니다.

Scikit-learn 라이브러리에서 제공하는 로지스틱회귀 모델을 초기화 하고 하이퍼파라미터 그리드를 설정하여 최적의 하이퍼파리미터를 찾기 위해 사용할 값들의 집합을 정의하였습니다.

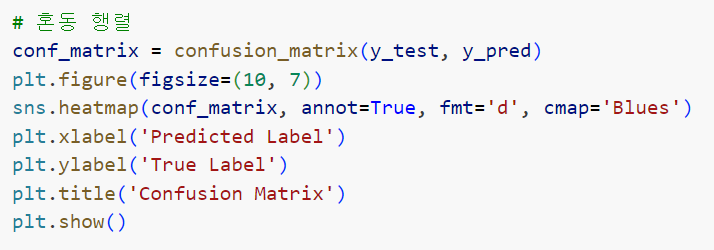
그리드 서치 클래스를 통해 지정된 하이퍼파라미터 그리드를 사용하여 교차 검증을 통해 최적의 하이퍼파리미터 조합을 찾았습니다.

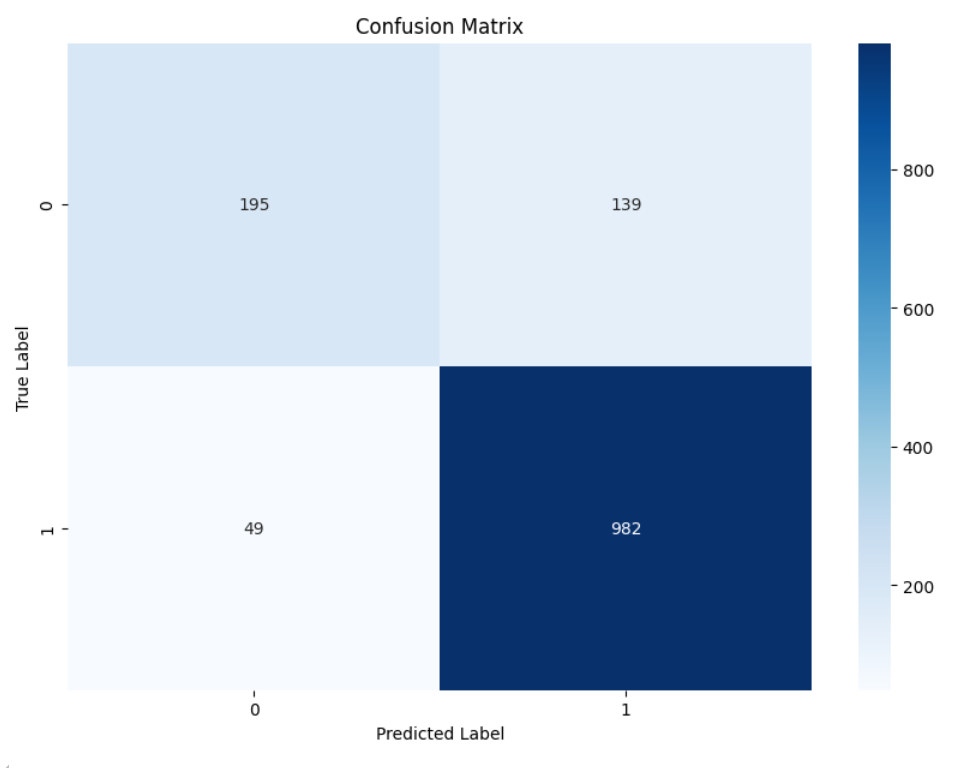
Best\_model 은 그리드 서치를 통해 찾은 최적의 하이퍼파리미터 조합을 사용한 최적의 모델입니다. 이 모델은 교차 검증 동안 가장 높은 정확도를 기록한 하이퍼파라미 조합으로 학습된 모델입니다.



이 부분에서는 정확도, 정밀도, 재현율, F1 점수, ROC-AUC를 계산하였습니다.

각 성능 지표를 계산한 후, 이를 출력하여 모델의 성능을 확인하였습니다. 정확도는 모델의 전반적인 성능을 나타내고, 정밀도와 재현율은 모델의 예측 오류 유형을 파악하기 위해 나타내었습니다. F1점수는 정밀도와 재현율 간의 균형을 제공하고 ROC-AUC는 모델의 분류 성능을 종합적을 평가하기 위해 나타내었습니다.



모델의 성능을 시각적으로 평가하기 위해 confusion\_matrix 함수를 사용하여 혼동 행렬을 생성하였습니다.

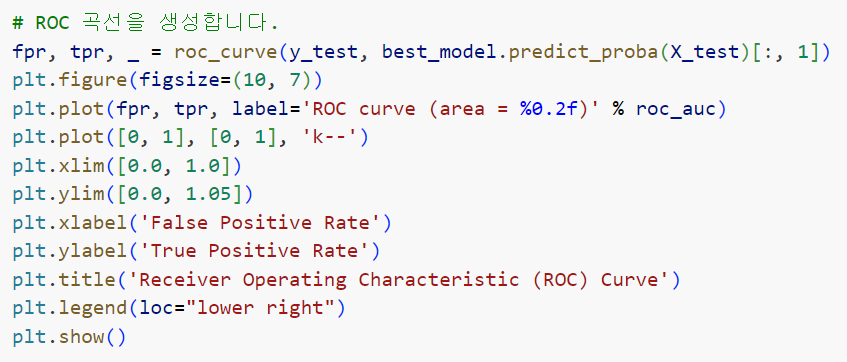
TN : 모델이 195개의 부정 리뷰를 정확히 부정으로 예측,

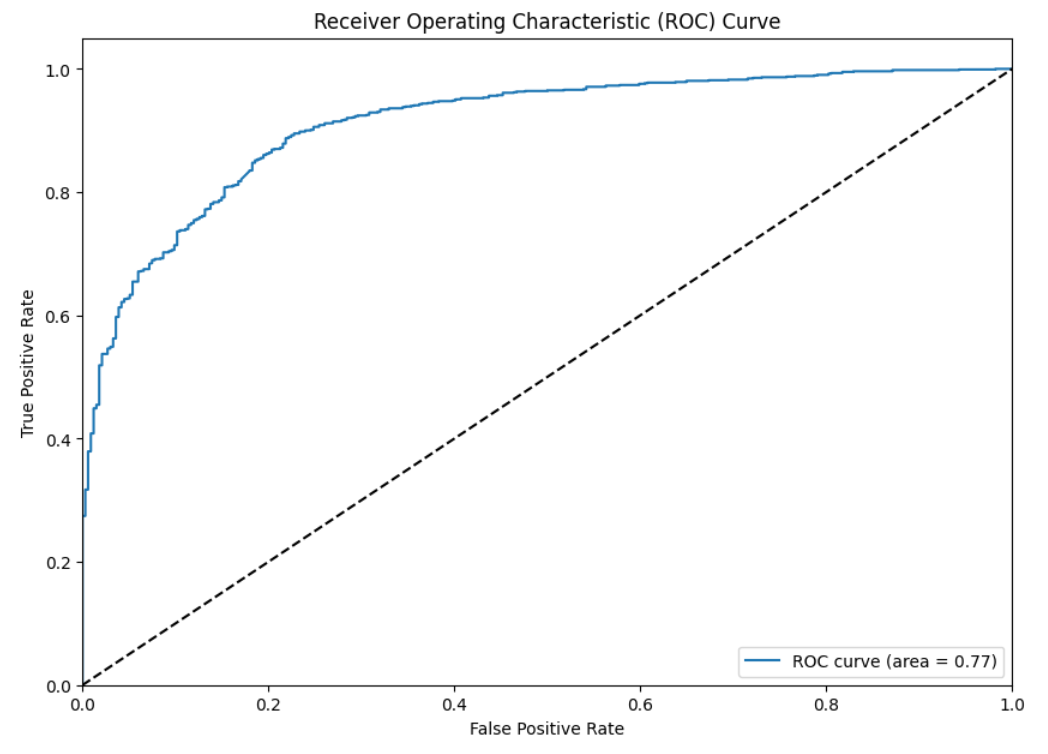
FP : 모델이 139개의 부정 리뷰를 긍정으로 잘못 예측,

FN : 모델이 49개의 긍정 리뷰를 부정으로 잘못 예측,

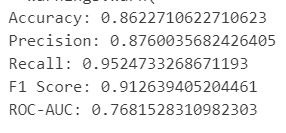
TP : 모델이 982개의 긍정 리뷰를 정확히 긍정으로 예측한 시각화 결과를 확인할 수 있었습니다.

따라서 이 모델은 높은 재현율(0.953)을 보여 긍정 리뷰를 잘 예측하고 있으며, 정밀도(0.876)도 비교적 높습니다. 그러나 139개의 부정 리뷰를 긍정으로 잘못 예측하는 문제가 있었습니다. 전반적으로, 모델은 긍정 리뷰를 매우 잘 예측하지만, 부정 리뷰에 대해서는 다소 오차가 있는 편으로 보여졌습니다.





이 부분에서는 최적의 모델을 사용하여 테스트 세트에 대한 ROC곡선을 생성하고 시각화하였습니다. ROC 곡선은 모델의 분류 성능을 다양한 임곗값에서 평가하며, AUC 값은 모델의 전반적인 성능을 나타냅니다. ROC 곡선과 AUC 값을 통해 모델의 강점과 약점을 보다 명확하게 이해하고, 모델의 성능을 종합적으로 평가할 수 있습니다.



**결론**

모델은 0.953의 높은 재현율을 보여 긍정 리뷰를 잘 예측하고 있으며, 정밀도도 0.876으로 비교적 높습니다. 하지만 139개의 부정 리뷰를 긍정으로 잘못 예측하는 문제가 있는 것이 확인되어 전반적으로 긍정 리뷰는 잘 예측하지만, 부정 리뷰에 대해서는 오차가 있는 편인 것을 알 수 있습니다. 또한 ROC 곡선이 대각선 기준선 위에 위치하며, 전체적으로 왼쪽 위로 향하고 있고, AUC 값은 0.77로, 모델이 상당히 양호한 성능을 가지고 있음을 알 수 있었습니다. 이는 무작위 예측(AUC = 0.5)보다 훨씬 우수하지만, 최적의 성능(AUC = 1)에는 다소 미치지 못하다는 것을 알 수 있었습니다.