기계학습 기말고사 대체과제 보고서

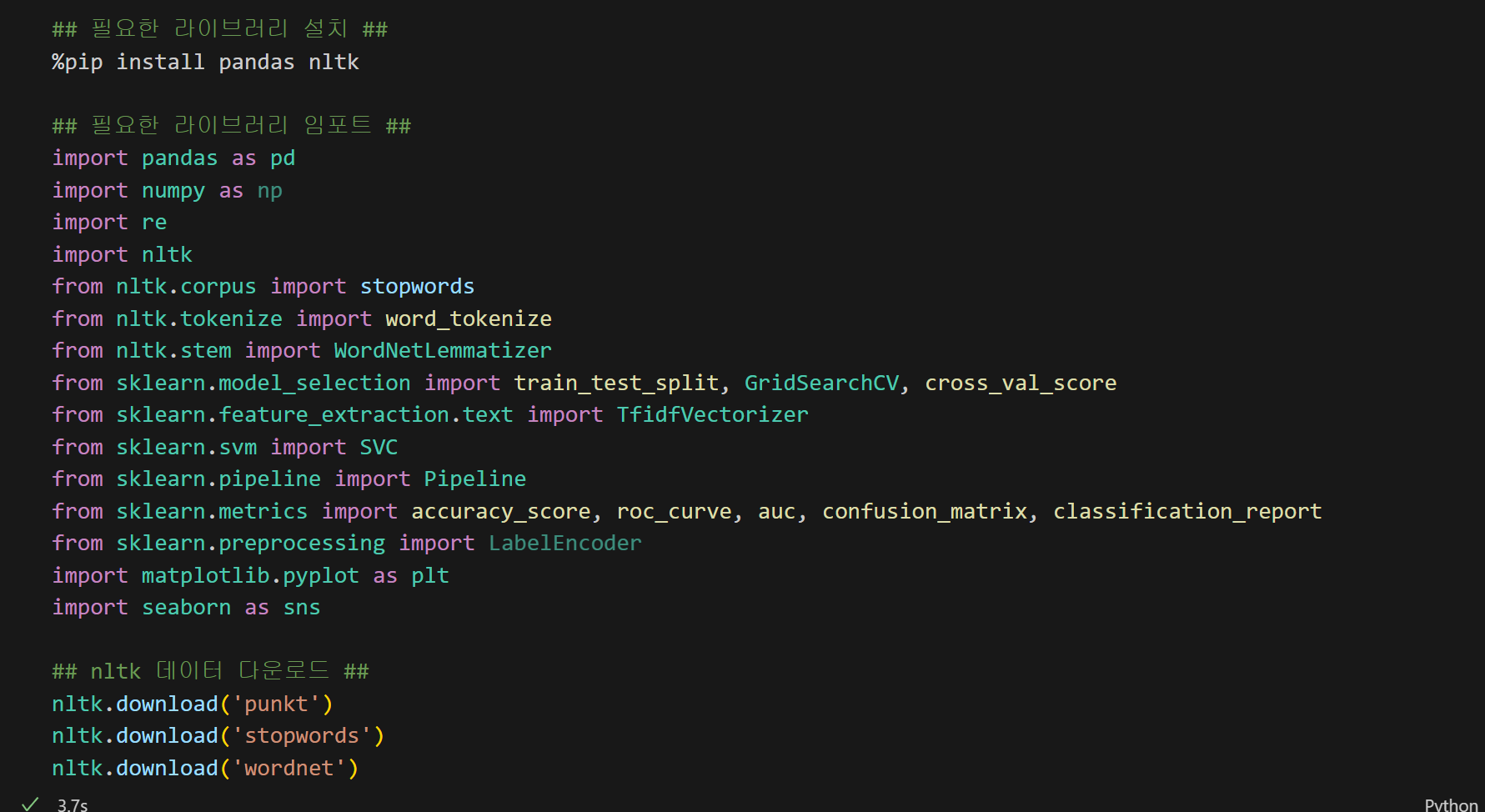
201920960 어형준

**1. 감성 분석의 기본 이해**

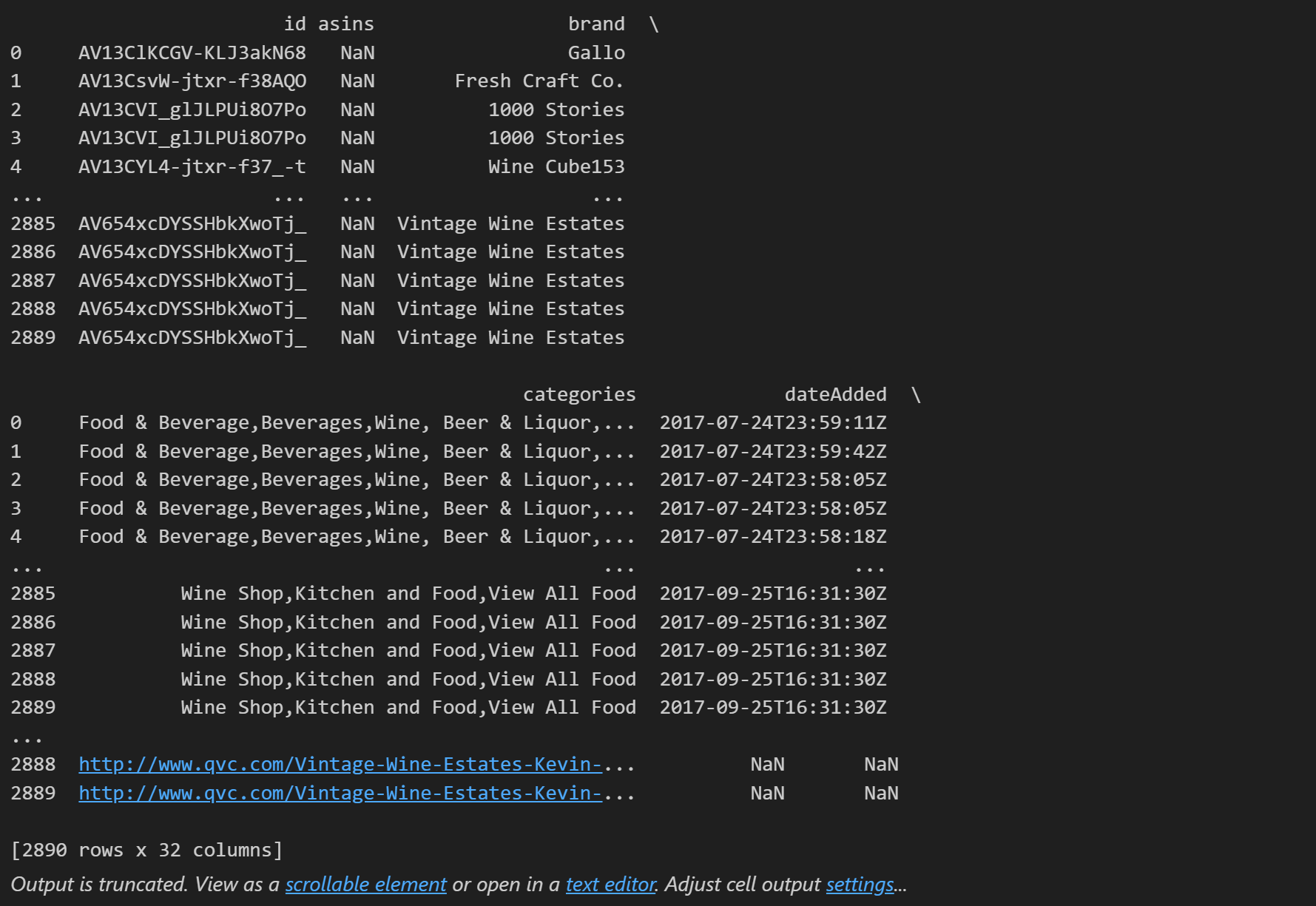
감성분석은 디지털 텍스트를 분석하여 메시지의 감정적 어조가 긍정적인지 부정적인지를 판단하는 것을 컴퓨터를 통해 분석하는 것입니다. 감성분석으로 wine\_review.csv 데이터에서 긍정 평가와 부정 평가를 평가할 수 있습니다.

**2. 데이터 전처리**

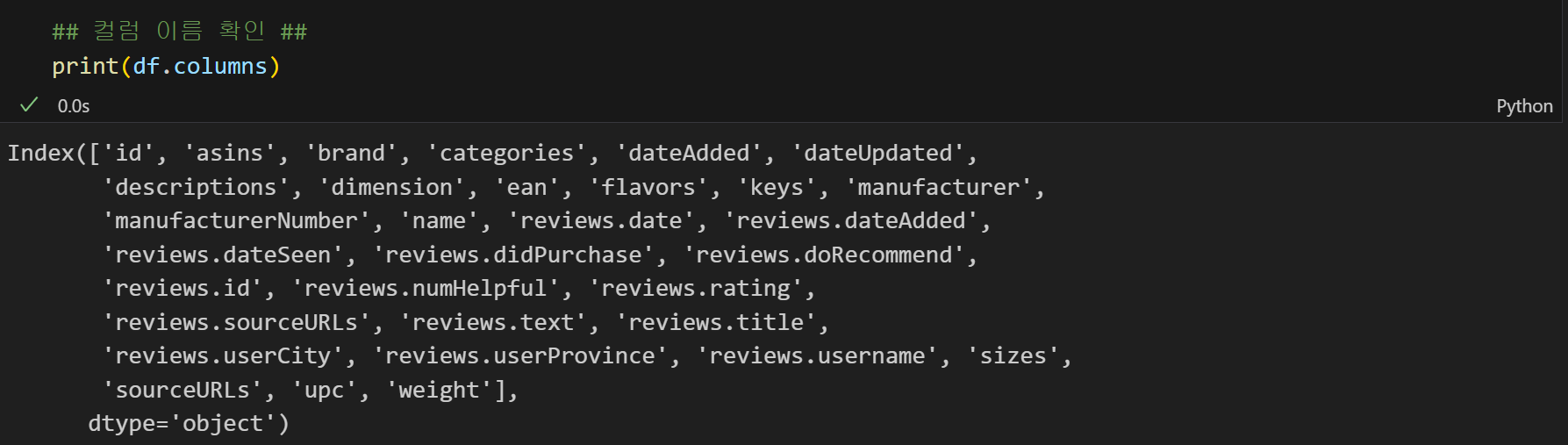
가장 먼저 필요한 라이브러리 설치 및 임포트를 하고, nltk 데이터를 다운받습니다.



다음으로 필요한 데이터를 로드하고, Print(df)로 불러온 데이터를 확인해줍니다.

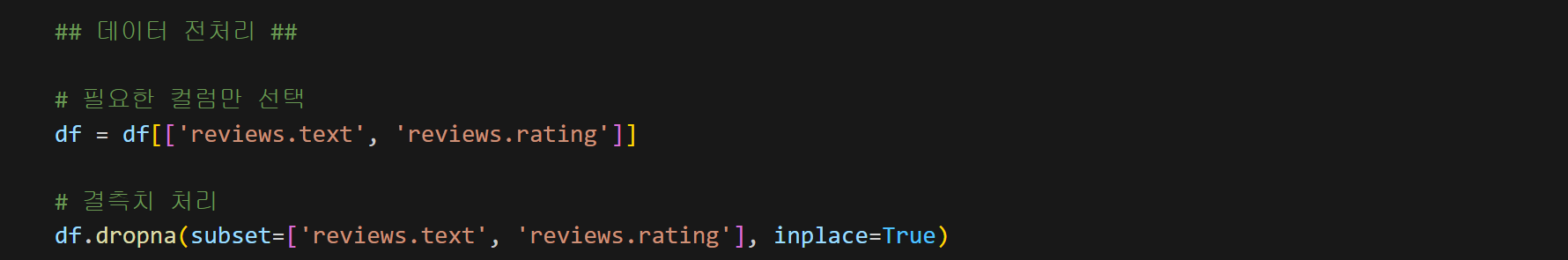
 

필요한 컬럼이 무엇인지 확인하기 위해 컬럼 이름을 확인해줍니다.

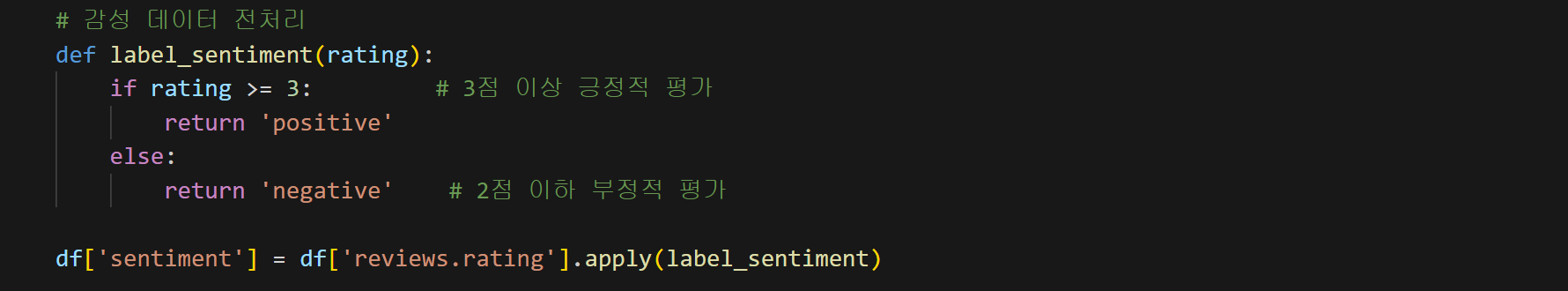


본격적인 데이터 전처리를 시작하겠습니다.

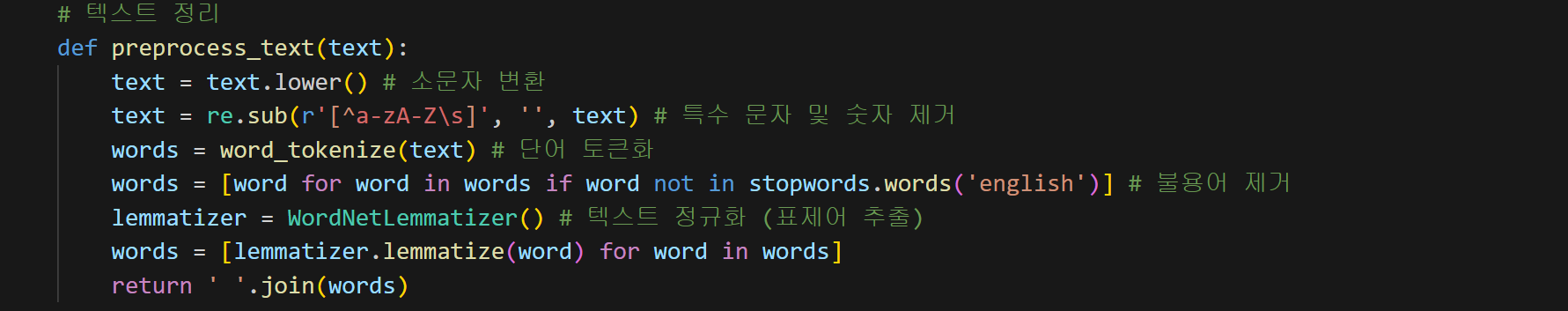
필요한 컬럼을 선택하고, 해당 컬럼에서 결측치를 처리해줍니다.



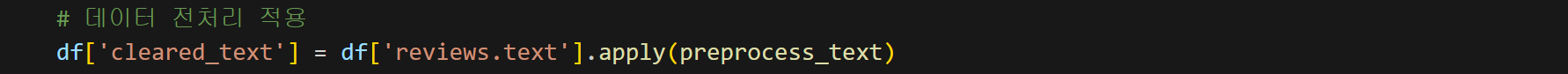
그리고 감성 데이터 전처리 단계에서 리뷰 점수 3점 이상은 긍적적 평가, 2점 이하는 부정적 평가로 감성 데이터를 전처리해줍니다. 해당 데이터를 ‘sentiment’ 데이터프레임으로 저장합니다.



이제 텍스트 정리를 해보겠습니다. 이 단계에서는 대문자를 소문자로 변환하고, 특수문자 및 숫자 제거, 단어 토큰화, 불용어 제거, 텍스트 정규화를 실시하였습니다.



이렇게 텍스트 정리가 완료된 데이터를 ‘cleared\_text’ 로 저장하였습니다.

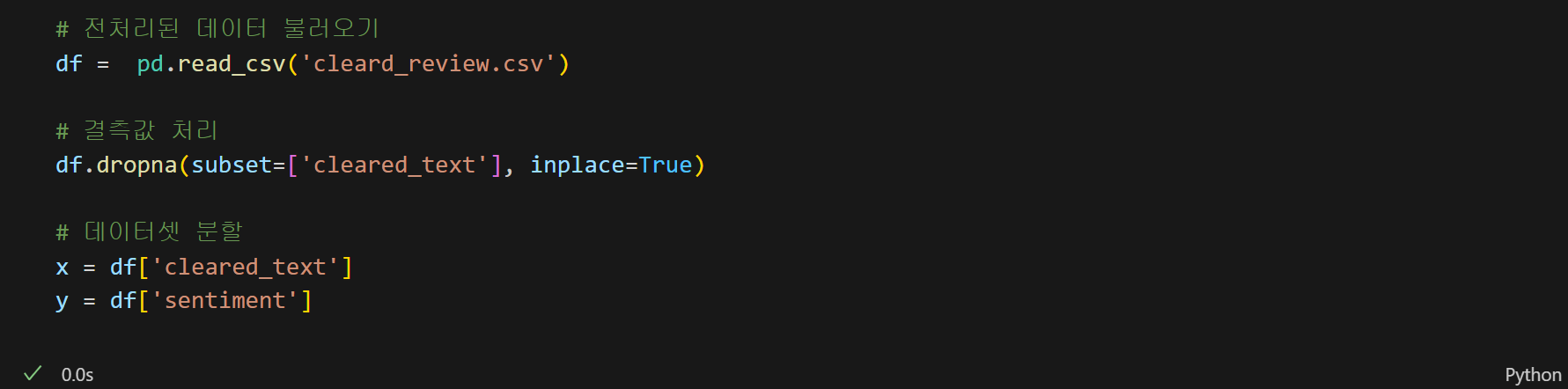


지금까지 만든 ‘sentiment’ 데이터와 ‘cleared\_text’ 데이터를 x와 y로 분할하고 전처리가 완료된 데이터를 ‘cleared\_review.csv’ 로 새로 저장합니다.

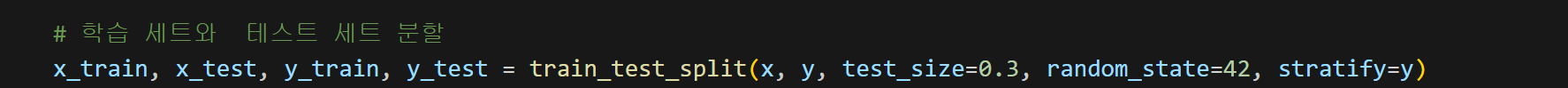


**3. 모델 선택 및 학습**

이제 모델 학습을 위한 전처리가 모두 끝났습니다. 여기서 한번 더 결측치를 제거해주고 데이터를 분할해줍니다.



train\_test\_split 함수를 사용해서 데이터를 학습 세트와 테스트 세트로 분할합니다.



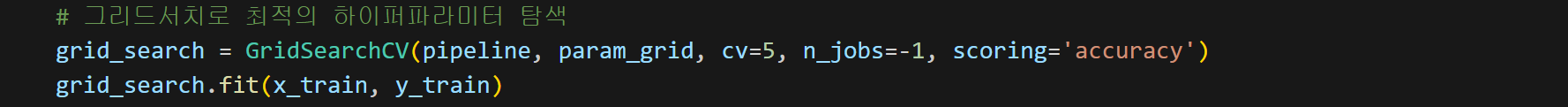
파이프라인 객체를 사용해서 전처리 단계와 모델 학습 단계를 이어줍니다.



다음은 하이퍼파라미터 튜닝입니다. param\_grid로 그리드서치를 통해 탐색하도록 하이퍼파라미터를 정의합니다..



이어서 그리드서치로 최적의 하이퍼파라미터를 탐색합니다.



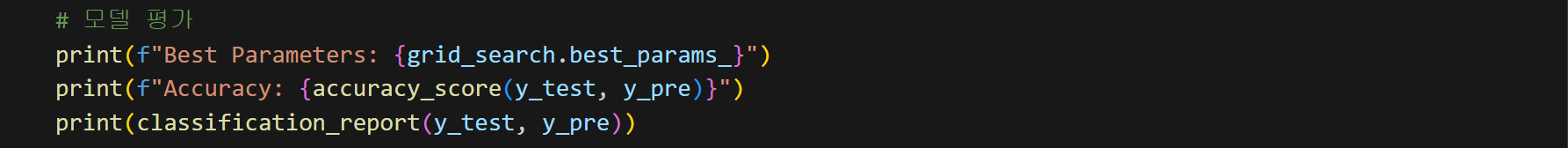
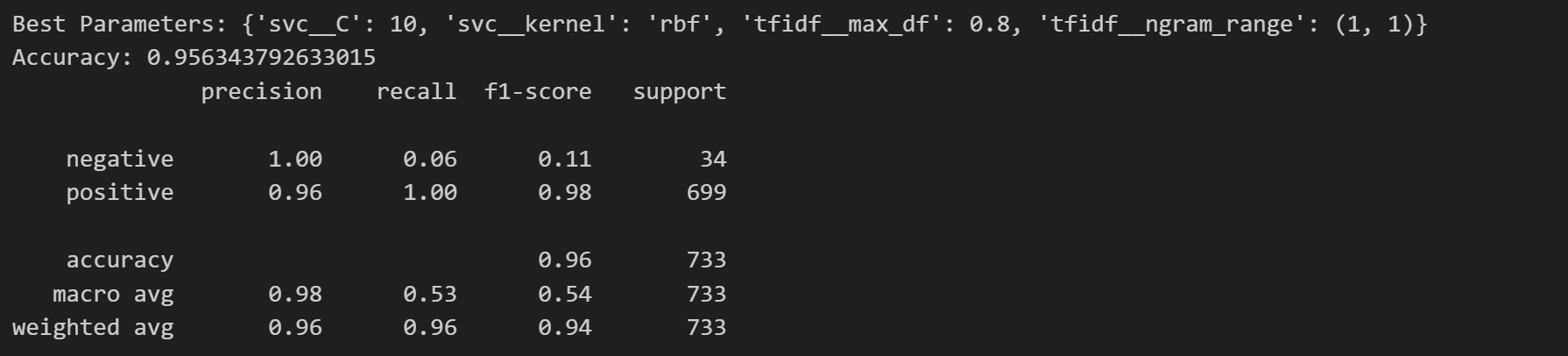
**4. 평가**

이제 모델 평가를 해보겠습니다.

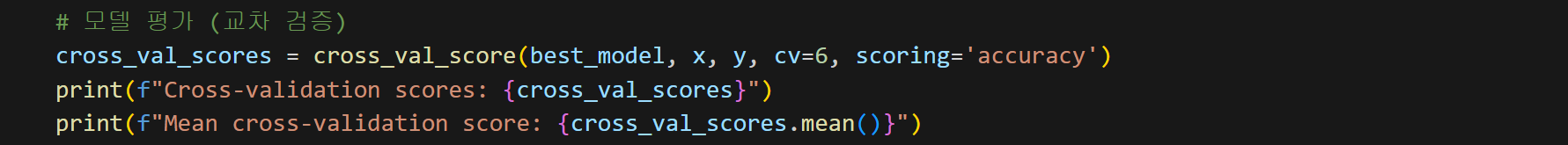
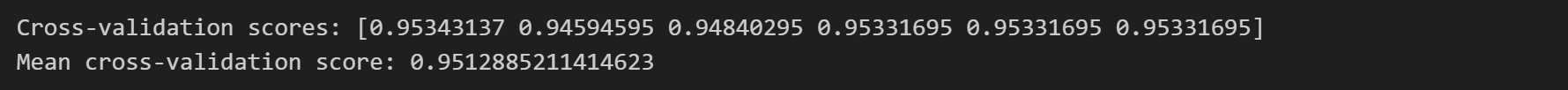
그리드서치로 찾은 최적의 하이퍼파라미터로 학습된 모델을 가져옵니다. 그리고 이 모델을 사용해 x\_test에 대한 예측 y\_pre를 만들어냅니다.



그리드서치로 찾은 최적의 하이퍼파라미터 값을 출력하고, y\_pre값과 y\_test 간의 정확도를 계산하고, 상세한 보고서를 출력합니다.

그리드서치로 찾은 최적의 하이퍼파라미터로 학습된 모델로 교차검증을 실시합니다. 이후 평균 확도를 출력합니다. 약 95.63%의 정확도를 보여줍니다. 또한 부정적 평가와 긍정적 평가의 예측 결과를 볼 수 있는데, 부정적 평가 예측의 값이 문제가 있어 보입니다. 하지만 전체적으로 보았을 때는 96%의 높은 정확성을 보이고 있습니다.

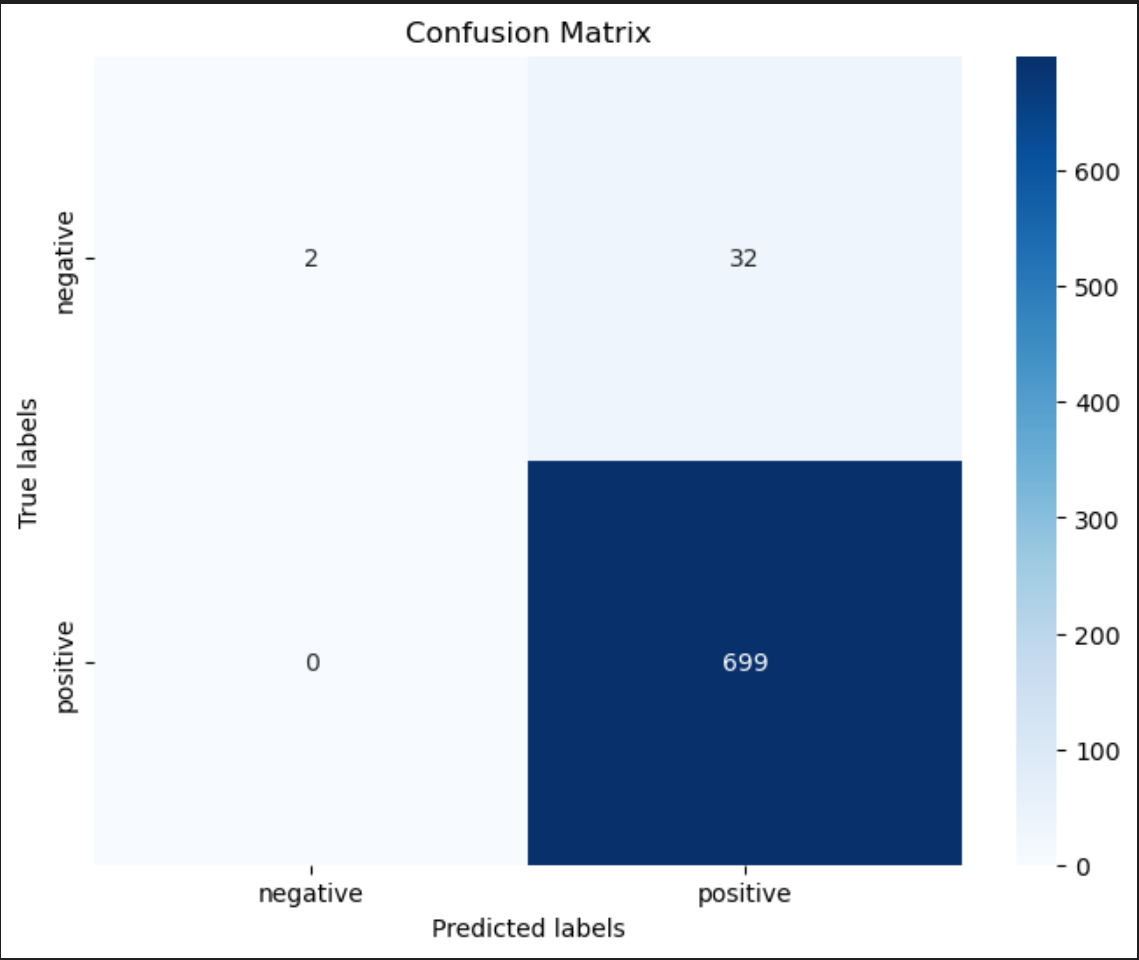
 

평균 정확도는 약 95.13%로 높은 정확성을 보여줍니다. 이는 모델이 정상적으로 작동하는 모습을 보여준다고 이해할 수 있습니다.

전체적으로 봤을 때는 정확도가 높았으나 부정적 평가 예측의 정확성이 떨어지는 것을 확인할 수 있습니다. 반면에 전체적으로 봤을 때에는 정확도가 높았습니다. 데이터를 열어보니 긍정적 평가가 압도적 비중을 차지하고 있는 것을 확인했습니다. 이러한 문제를 해결하려면 데이터를 추가하는 방법이 있겠습니다.

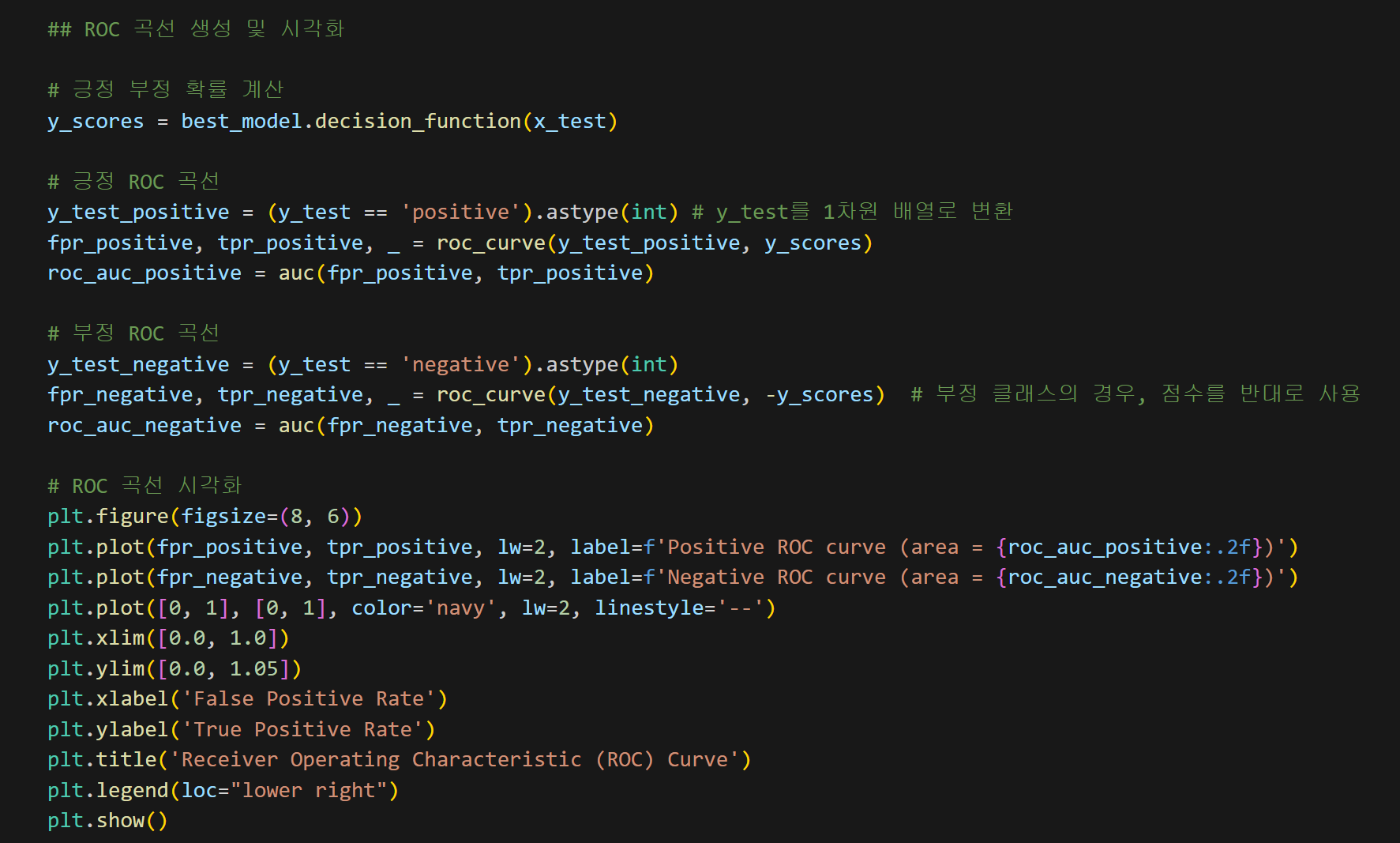
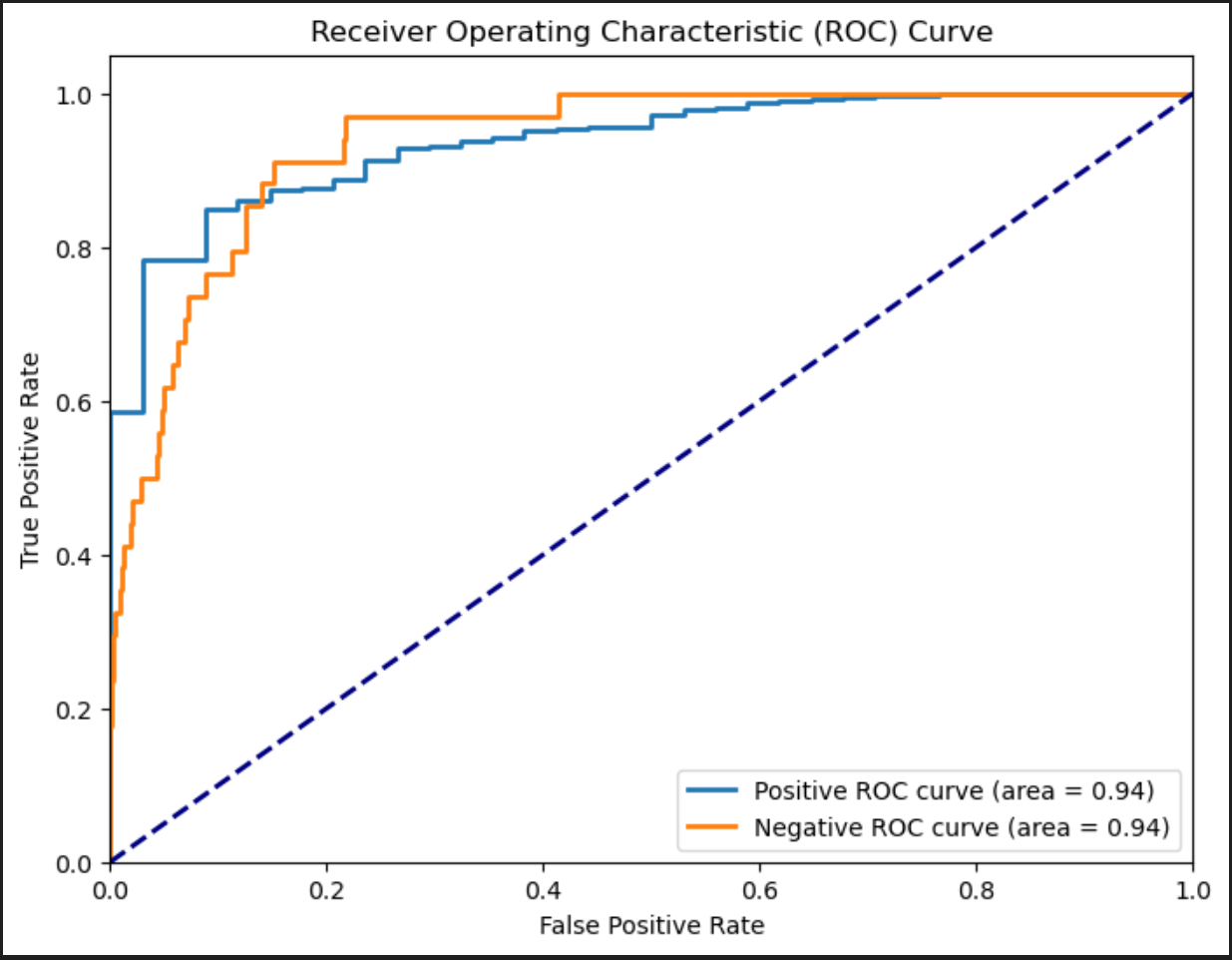
**5. 시각화 및 해석**

먼저 혼동행렬로 시각화를 하였습니다.

해당 결과를 통해 분류가 잘되었지만 긍정적 반응을 예측하는 쪽으로 몰려 있는 것을 볼 수 있습니다. 실제 데이터를 살펴보니 긍정적인 리뷰가 많아서 발생한 현상으로 추측됩니다. 더 다양한 반응의 리뷰가 섞여 있었다면 전체적으로 분포되어 있는 모습을 예측해 볼 수 있습니다.

마지막으로 ROC 곡선입니다.

두 곡선 모두 AUC가 0.94로 긍정적 평가 예측과 부정적 평가 예측 모두 성능이 좋다는 것을 알 수 있습니다.