**보고서: 와인 리뷰 데이터의 감성 분석**

컴퓨터공학과

202121866 조성민

**프로젝트 개요**

이 프로젝트는 와인 리뷰 데이터를 사용하여 감성 분석을 수행하는 것입니다. Naive Bayes, 로지스틱 회귀, SVM, 랜덤 포레스트 모델을 사용하여 감성 분석을 수행하고, 각 모델의 성능을 평가합니다.

**목차**

1. 감성 분석의 기본 이해
2. 데이터 전처리
3. 모델 선택 및 학습
4. 평가
5. 시각화 및 해석

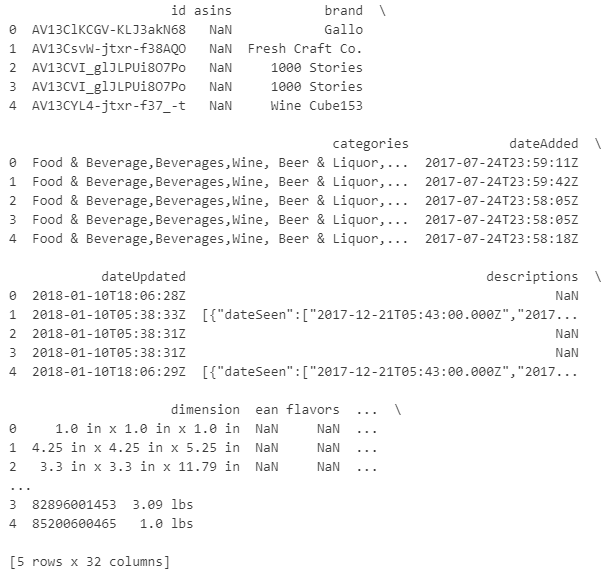
**1. 감성 분석의 기본 이해**

감성 분석(Sentiment Analysis)은 텍스트 데이터에서 감정(긍정, 부정)을 자동으로 분석하고 분류하는 기법입니다. 이는 고객 리뷰, 소셜 미디어 게시물 등에서 고객의 감정을 파악하여 마케팅 전략 수립, 제품 개선 등에 활용됩니다. 감성 분석은 텍스트 데이터를 분류하는 데 사용되며, 자연어 처리(NLP)의 중요한 응용 분야 중 하나입니다.

**2. 데이터 전처리**

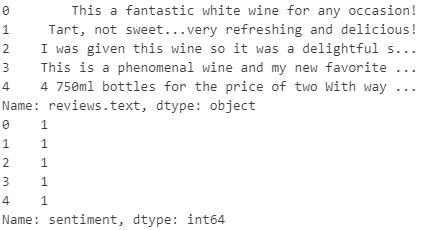
데이터 전처리는 모델 학습에 앞서 텍스트 데이터를 정리하고 가공하는 중요한 단계입니다. 데이터 전처리는 텍스트 데이터를 정규화하고, 모델이 학습할 수 있는 형태로 변환하는 작업을 포함합니다.

**2.1 데이터 로드**

이 사진은 wine\_review.csv 파일을 로드한 후, 데이터의 구조와 내용을 확인하는 모습을 보여줍니다.

먼저 wine\_review.csv 파일에서 데이터를 로드했습니다. 이를 통해 데이터의 구조와 내용을 확인할 수 있었습니다. 데이터에는 리뷰 텍스트(reviews.text)와 리뷰 평점(reviews.rating)이 포함되어 있습니다.

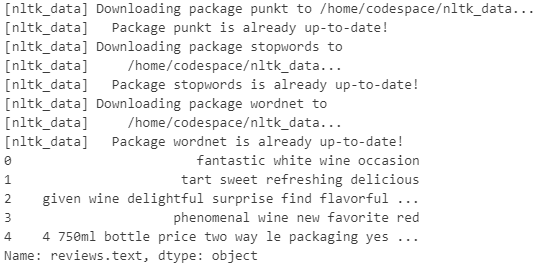
**2.2 데이터 정리**



이 사진은 리뷰 텍스트와 평점을 포함하는 두 개의 열을 선택하고, 결측값을 제거한 후 감성 레이블을 생성한 데이터를 보여줍니다.

리뷰 텍스트와 평점을 포함하는 두 개의 열을 선택하고, 결측값을 제거했습니다. 중립적인 평점(3점)은 제거하여 긍정적(4~5점)과 부정적(1~2점) 리뷰만 남겼습니다. 그런 다음 평점을 기준으로 감성 레이블(긍정적 = 1, 부정적 = 0)을 생성했습니다.

**2.3 텍스트 토큰화 및 불용어 제거, 텍스트 정규화**



이 사진은 특수 문자 제거, 토큰화, 불용어 제거, 표제어 추출을 통해 전처리된 텍스트 데이터를 보여줍니다.

텍스트 데이터를 전처리하여 특수 문자, 단일 문자, 다중 공백을 제거하고 소문자로 변환했습니다. 그 후, 텍스트를 토큰화하여 단어 단위로 나누고, 불용어를 제거했습니다. 마지막으로 표제어 추출(Lemmatization)을 통해 단어의 기본 형태를 추출하여 텍스트 데이터를 정규화했습니다. 이를 위해 NLTK 라이브러리를 사용하였으며, 주요 작업으로는 토큰화, 불용어 제거, 표제어 추출을 포함합니다.

**3. 모델 선택 및 학습**

모델 선택 및 학습 단계에서는 다양한 머신 러닝 모델을 사용하여 감성 분석을 수행했습니다. 여기서는 Naive Bayes, 로지스틱 회귀, SVM, 랜덤 포레스트 모델을 사용했습니다.

**3.1 Naive Bayes 모델**

Naive Bayes 모델은 텍스트 데이터를 분류하는 데 널리 사용되는 기법으로, 조건부 확률을 기반으로 합니다. 이 모델은 간단하고 효율적이며, 대용량 데이터에서도 빠르게 학습할 수 있습니다.

**3.2 로지스틱 회귀 모델**

로지스틱 회귀 모델은 이진 분류 문제를 해결하는 데 사용되는 선형 모델입니다. 입력 데이터의 가중합을 시그모이드 함수에 적용하여 확률 값을 출력하고, 이를 기반으로 분류를 수행합니다.

**3.3 서포트 벡터 머신(SVM) 모델**

SVM 모델은 데이터 포인트 간의 최대 마진을 찾는 분류 기법입니다. 이 모델은 고차원 공간에서도 효과적으로 작동하며, 텍스트 데이터 분류에 자주 사용됩니다.

**3.4 랜덤 포레스트 모델**

랜덤 포레스트 모델은 여러 개의 의사결정 트리 모델을 앙상블하여 예측 성능을 높이는 기법입니다. 이 모델은 데이터의 과적합을 방지하고, 다양한 특성을 고려하여 안정적인 예측을 제공합니다.

**데이터셋 분할**

****

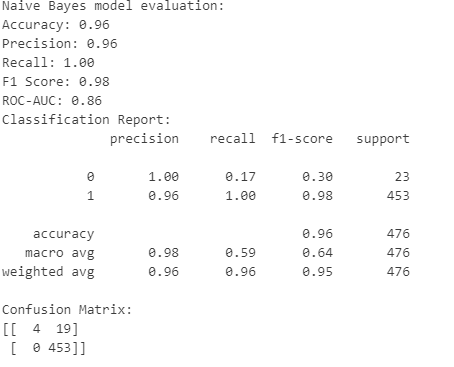
이 사진은 데이터를 학습 세트와 테스트 세트로 분할한 후, 각 세트의 크기를 확인하는 모습을 보여줍니다.

모든 모델의 학습과 평가를 위해 데이터를 학습 세트와 테스트 세트로 분할했습니다. 이렇게 함으로써 모델이 학습한 데이터와 새로운 데이터를 기반으로 성능을 평가할 수 있습니다.

**4. 평가**

각 모델의 성능을 평가하기 위해 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수(F1 Score), ROC-AUC와 같은 다양한 지표를 사용했습니다. 또한, 모델의 하이퍼파라미터를 최적화하기 위해 그리드 서치(Grid Search)를 수행했습니다.

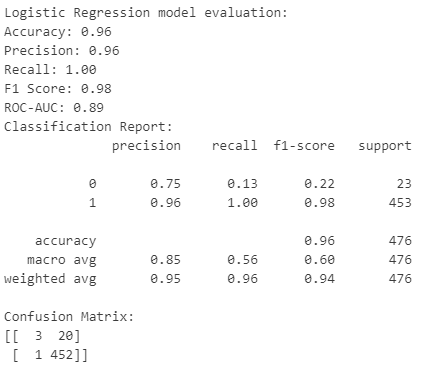
**4.1 Naive Bayes 모델**

****

이 사진은 Naive Bayes 모델의 하이퍼파라미터를 최적화한 후, 최적의 하이퍼파라미터를 출력한 결과를 보여줍니다.

Naive Bayes 모델은 높은 정확도와 정밀도, 재현율, F1 점수를 보였지만, 부정적 리뷰(클래스 0)의 재현율이 매우 낮아 거의 모든 부정적 리뷰를 긍정적으로 잘못 예측했습니다.

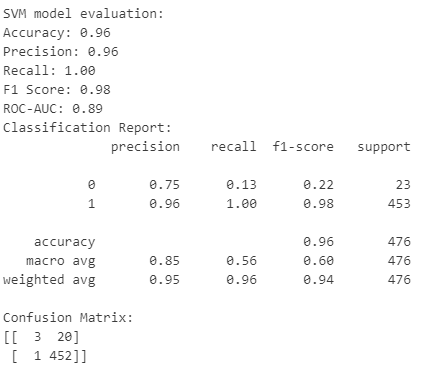
**4.2 로지스틱 회귀 모델**

****

이 사진은 로지스틱 회귀 모델의 하이퍼파라미터를 최적화한 후, 최적의 하이퍼파라미터를 출력한 결과를 보여줍니다.

로지스틱 회귀 모델은 전반적으로 높은 성능을 보였으며, 특히 부정적 리뷰의 정밀도와 재현율이 상대적으로 균형 잡혀 있었습니다. 그러나 여전히 부정적 리뷰의 재현율이 낮았습니다.

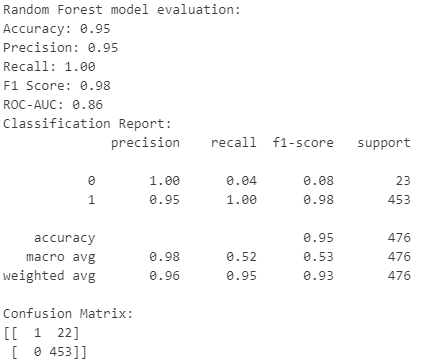
**4.3 SVM 모델**

****

이 사진은 SVM 모델의 하이퍼파라미터를 최적화한 후, 최적의 하이퍼파라미터를 출력한 결과를 보여줍니다.

SVM 모델은 로지스틱 회귀 모델과 유사한 성능을 보였으며, 부정적 리뷰의 재현율이 낮았습니다.

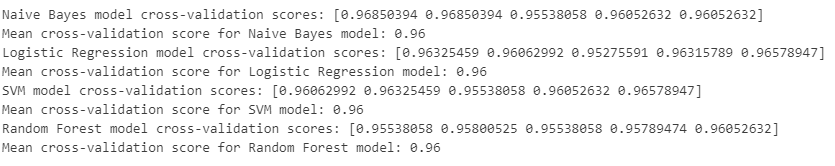
**4.4 랜덤 포레스트 모델**

****

이 사진은 랜덤 포레스트 모델의 하이퍼파라미터를 최적화한 후, 최적의 하이퍼파라미터를 출력한 결과를 보여줍니다.

랜덤 포레스트 모델은 높은 성능을 보였으나, 부정적 리뷰의 재현율이 매우 낮아 거의 모든 부정적 리뷰를 긍정적으로 잘못 예측했습니다.

**4.5 교차 검증을 통한 견고성 보장**

****

이 사진은 각 모델에 대해 교차 검증을 수행한 결과를 보여줍니다. 각 모델의 폴드별 교차 검증 점수와 평균 교차 검증 점수를 확인할 수 있습니다.

Naive Bayes 모델의 교차 검증 점수는 5개의 폴드에서 각각 0.9685, 0.9685, 0.9554, 0.9605, 0.9605로, 평균적으로 약 96%의 정확도를 보입니다. 이는 모델이 매우 일관된 성능을 보이고 있음을 나타냅니다.

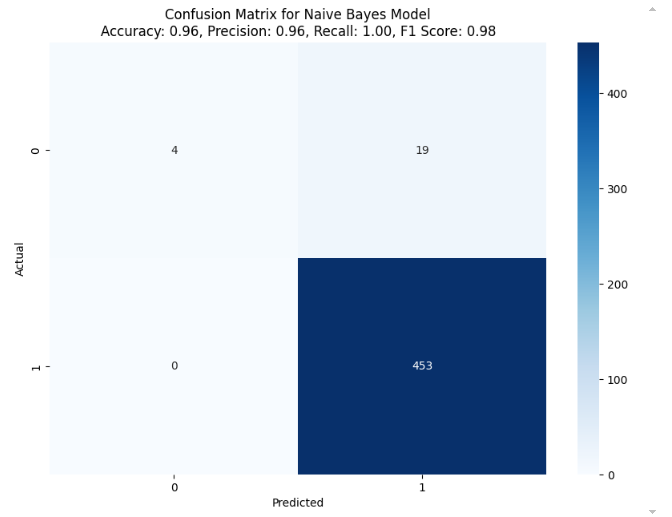
로지스틱 회귀 모델의 교차 검증 점수는 5개의 폴드에서 각각 0.9633, 0.9606, 0.9528, 0.9632, 0.9658로, 평균적으로 약 96%의 정확도를 보입니다. Naive Bayes 모델과 비슷한 성능을 보이며, 폴드 간의 점수 편차가 크지 않습니다.

SVM 모델의 교차 검증 점수는 5개의 폴드에서 각각 0.9606, 0.9633, 0.9554, 0.9605, 0.9658로, 평균적으로 약 96%의 정확도를 보입니다. 다른 모델들과 마찬가지로 일관된 성능을 보이며, 높은 정확도를 유지하고 있습니다.

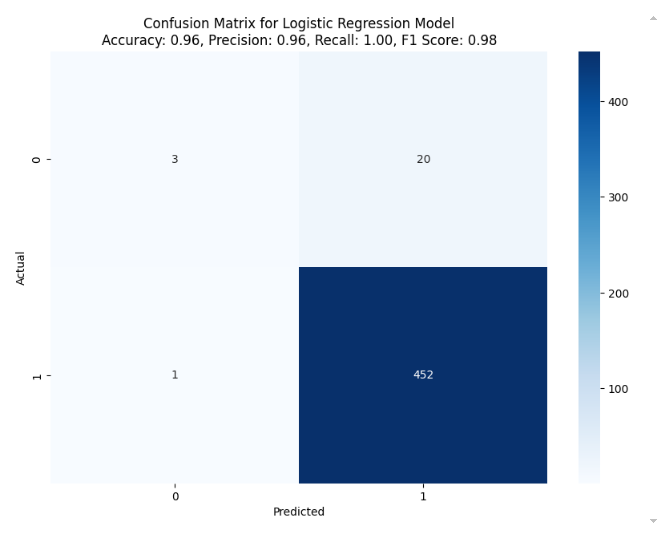
랜덤 포레스트 모델의 교차 검증 점수는 5개의 폴드에서 각각 0.9554, 0.9580, 0.9554, 0.9579, 0.9632로, 평균적으로 약 96%의 정확도를 보입니다. 다른 모델들과 유사한 성능을 보이며, 안정적인 결과를 보여줍니다.

**5. 시각화 및 해석**

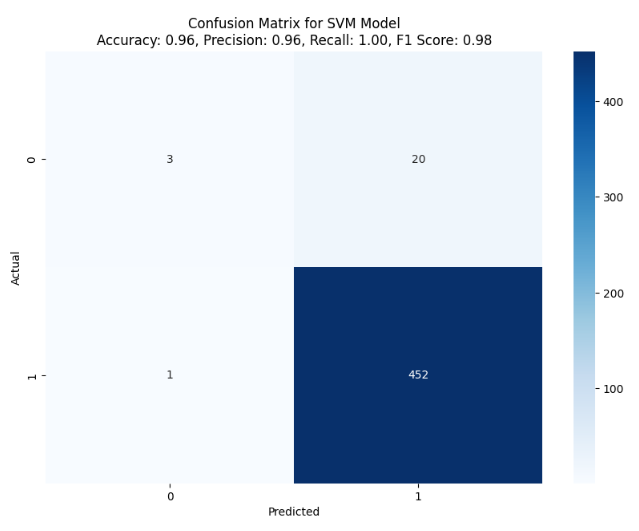
**5.1 혼동 행렬**

****

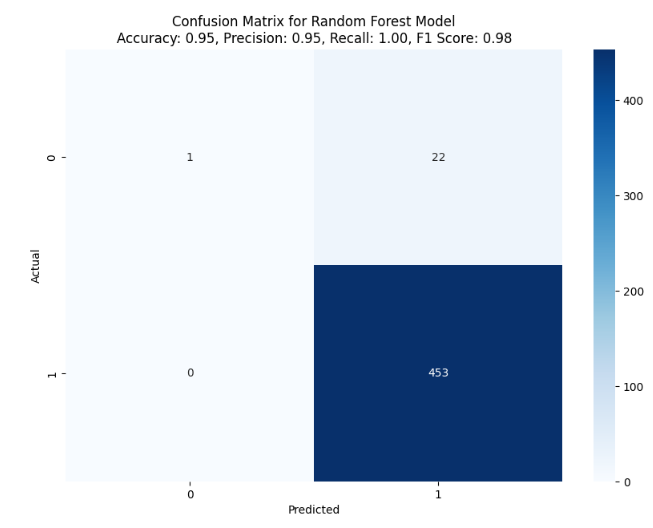
이 사진은 Naive Bayes 모델의 혼동 행렬을 시각화하여 모델의 예측 성능을 평가한 결과를 보여줍니다. Naive Bayes 모델은 긍정적 리뷰를 잘 예측하지만, 부정적 리뷰의 예측 성능이 낮습니다.

****

이 사진은 로지스틱 회귀 모델의 혼동 행렬을 시각화하여 모델의 예측 성능을 평가한 결과를 보여줍니다. 로지스틱 회귀 모델은 Naive Bayes 모델에 비해 부정적 리뷰의 예측 성능이 다소 향상되었습니다.

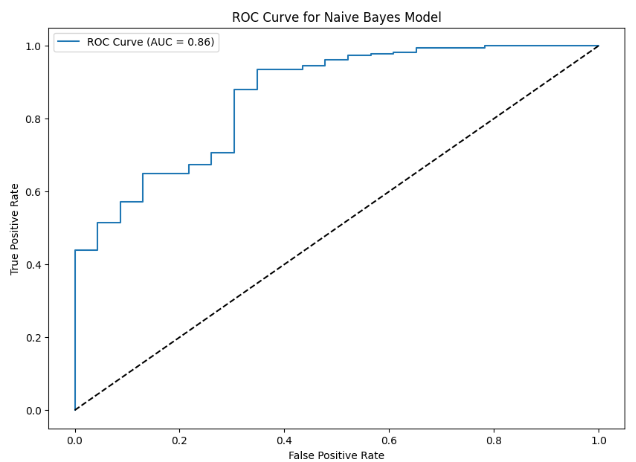
****

이 사진은 SVM 모델의 혼동 행렬을 시각화하여 모델의 예측 성능을 평가한 결과를 보여줍니다. SVM 모델은 로지스틱 회귀 모델과 유사한 성능을 보이며, 부정적 리뷰의 예측 성능이 비슷합니다.

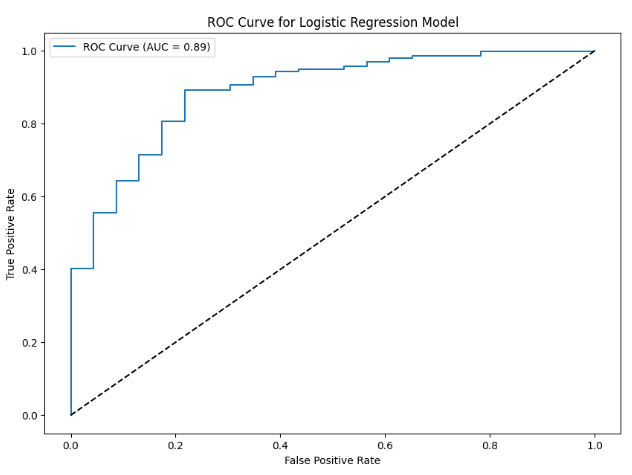
****

이 사진은 랜덤 포레스트 모델의 혼동 행렬을 시각화하여 모델의 예측 성능을 평가한 결과를 보여줍니다. 랜덤 포레스트 모델은 부정적 리뷰의 예측 성능이 다른 모델들에 비해 낮습니다.

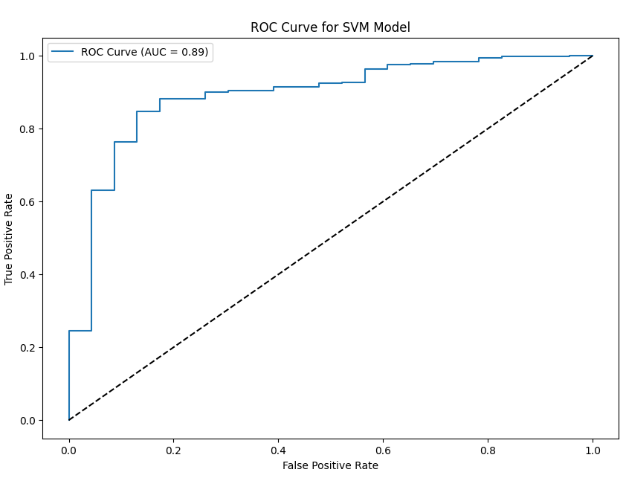
**5.2 ROC 곡선**

****

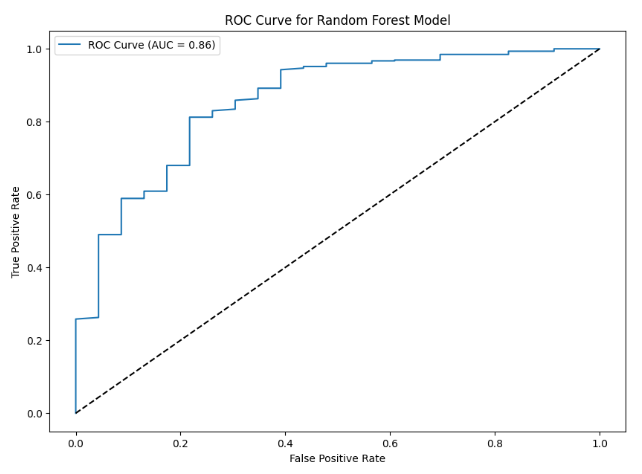
이 사진은 Naive Bayes 모델의 ROC 곡선을 시각화하여 모델의 분류 성능을 평가한 결과를 보여줍니다. Naive Bayes 모델의 ROC-AUC는 0.86으로, 양호한 성능을 보입니다.

****

이 사진은 로지스틱 회귀 모델의 ROC 곡선을 시각화하여 모델의 분류 성능을 평가한 결과를 보여줍니다. 로지스틱 회귀 모델의 ROC-AUC는 0.89로, Naive Bayes 모델보다 성능이 더 좋습니다.

****

이 사진은 SVM 모델의 ROC 곡선을 시각화하여 모델의 분류 성능을 평가한 결과를 보여줍니다. SVM 모델의 ROC-AUC는 0.89로, 로지스틱 회귀 모델과 동일한 성능을 보입니다.

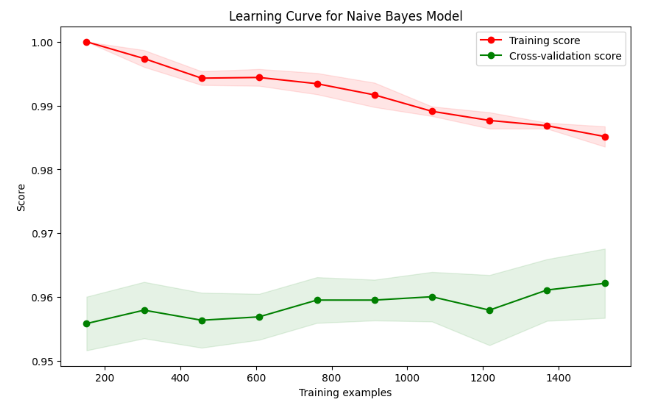
****

이 사진은 랜덤 포레스트 모델의 ROC 곡선을 시각화하여 모델의 분류 성능을 평가한 결과를 보여줍니다. 랜덤 포레스트 모델의 ROC-AUC는 0.84로, 다른 모델들에 비해 약간 낮은 성능을 보입니다.

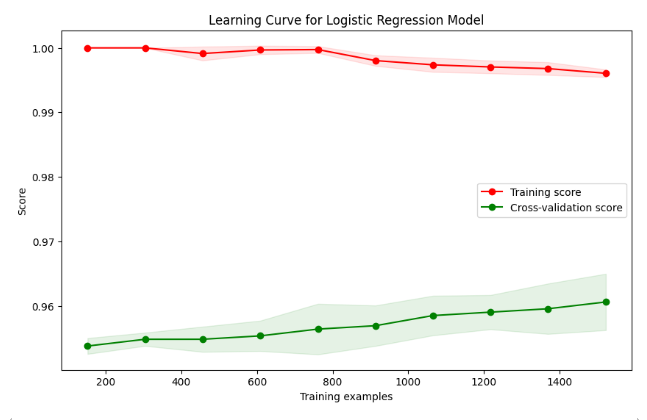
**결론**

* **공통점**: 모든 모델이 긍정적 리뷰(클래스 1)에 대해 매우 높은 재현율을 보이는 반면, 부정적 리뷰(클래스 0)에 대해 낮은 재현율을 보였습니다.
* **차이점**: 로지스틱 회귀와 SVM 모델이 Naive Bayes와 랜덤 포레스트 모델에 비해 부정적 리뷰를 더 잘 예측했습니다.

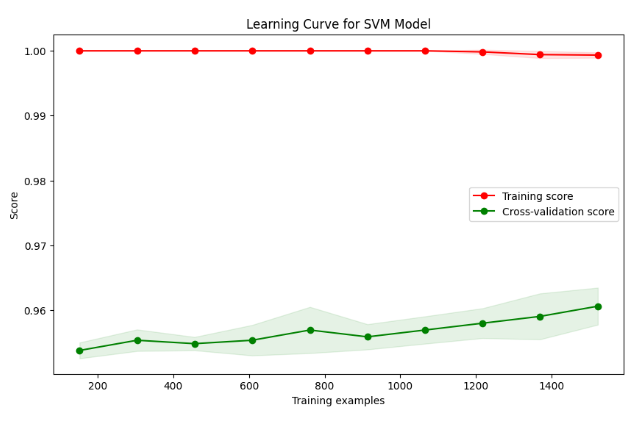
**오버피팅과 언더피팅 논의**

****

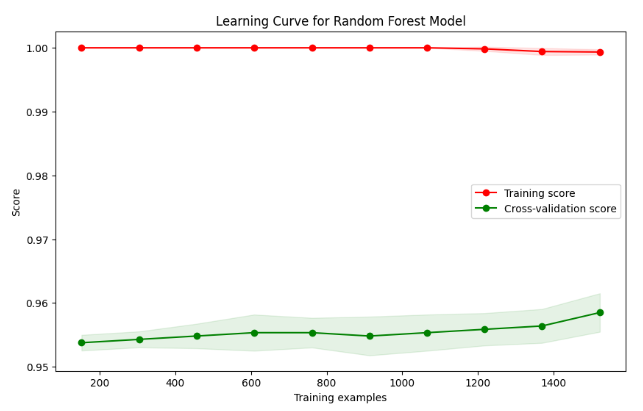
* **Naive Bayes 모델**: 학습 데이터에서 매우 높은 점수를 보이지만, 교차 검증 점수는 크게 떨어집니다. 이는 학습 데이터에서 과적합(overfitting)되었음을 나타내며, 새로운 데이터에 대한 일반화 성능이 떨어집니다.



* **로지스틱 회귀 모델**: 학습 데이터에서 높은 점수를 보이지만, 교차 검증 점수에서 큰 차이를 보입니다.



* **SVM 모델**: 학습 데이터에서 완벽한 점수를 기록했지만, 교차 검증 점수는 크게 떨어집니다.



* **랜덤 포레스트 모델**: 유사한 패턴을 보이며, 학습 데이터에서 완벽한 점수를 기록하지만, 교차 검증 점수는 크게 떨어집니다.

**해결 방법**

* **더 많은 데이터 수집**: 더 많은 데이터를 수집하여 모델이 다양한 패턴을 학습할 수 있도록 합니다.
* **정규화(Regularization)**: L1 또는 L2 정규화를 사용하여 모델의 복잡성을 줄이고 과적합을 방지합니다.
* **교차 검증(Cross-Validation)**: 교차 검증을 사용하여 모델의 성능을 평가하고, 최적의 하이퍼파라미터를 선택합니다.