|  |
| --- |
| 기말고사 |
| 과제 보고서 |
| 기계학습 |

|  |
| --- |
| 준영 신  23.3.13 |

이름: 신준영

학과: 소프트웨어전공

학번: 201920950

**1. 데이터 구조 및 분석**

데이터 이름: amazon\_uk\_shoes\_products\_dataset\_2021\_12.csv

컬럼 리스트: url, product\_name, reviewer\_name, review\_title, review\_text,

review\_rating, verified\_purchase, review\_date, helpful\_count,

uniq\_id, scraped\_at

URL 컬럼: 리뷰 대상 상품을 구매하기 위한 페이지

**Product\_name 컬럼**: 리뷰 대상 상품의 영문 이름

**reviewer\_name, review\_rating, review\_date 컬럼**: 리뷰 작성자의 이름, 리뷰 별점을 의미한다.

**review\_title, review\_text, 컬럼**: 각각 리뷰 이름, 리뷰 본문, 리뷰 본문을 의미한다. 이 칼럼의 데이터에는 주의해야할 특성이 있다.

1. 작성된 언어가 영어, 일본어, 독일어, 이탈리아어 등 다양한 언어로 구성되어 있다. 감성분석을 위해 사용되는 vader 라이브러리는 영어만 해석이 가능하므로 주의하도록 하자.

2. 리뷰 본문을 살펴보면 이모티콘과 같은 특수문자를 포함하고 있다. 사용하는 라이브러리에 따라 주의해줄 필요가 있지만 이번 프로젝트에서 사용된 라이브러리에서는 모두 처리가 가능 범위에 포함된다.

3. 리뷰 본문을 살펴보면 **'The media could not be loaded.\n \n\n\n\n\n \xc2\xa0 '** 와 같은 문자열이 포함된 리뷰를 확인 할 수 있다. 이와 같은 경우 리뷰에 이미지 파일이 포함된 경우로 추측되어, csv파일 작성시 이미지 파일을 이러한 문구로 대체한 것으로 보인다. 이와 같은 경우 학습에 악영향을 미칠 수 있으므로 제거해 주어야 한다.

4. 내용이 NaN으로 결측값이 존재한다. 적은 수이므로 데이터 정제시 제거해 준다.

5. \xc2\xa0와 같은 문자열은 html의 공백 문자로 일반적인 라이브러리에는 오동작을 하게되는 경우가 있으므로 반드시 제거되어야 한다.

**verified\_purchase 칼럼**: 이 칼럼은 리뷰 작성자가 실제로 상품을 구매했는지 확인할 수 있는 칼럼으로 True일 경우 실제로 구매한 경우이다. 다른 사이트에서의 구매, 할인 이벤트 중의 구매자 같은 경우 False로 표시되어 리뷰 진정성을 평가하기 위한 좋은 지표이다.

**helpful\_count 칼럼**: 리뷰에 대한 추천수로 양의 정수 값을 가진다. 추천수가 0명인 경우 NaN으로 작성되어 있어. 주의가 필요하다.

**uniq\_id, scraped\_at 칼럼:** 데이터에 부여된 index와 데이터가 생성된 시간을 의미한다.

**2. 데이터 정제**

1. verified\_purchase가 False인 리뷰는 제거한다.

2. 똑 같은 리뷰를 반복하여 작성되는 경우가 확인되나, 이런 경우 good, exellent와 같이 1,2단어로 작성된 경우로 분석 결과에 큰 영향을 미칠것으로 평과되지 않아 데이터를 보존하였다.

3. location에 따라 다양한 언어로 작성되어 있으므로 처리 전에 번역되어야 한다. 번역은 DeepLApi와 Langid 라이브러리를 사용하여 진행하였다. 코드를 다시 실행하는 경우 Api 비용이 부담되게 되므로 번역된 기록이 있는 경우 예전 번역을 불러온다. 번역 기록은 translation\_progress.pkl파일에 저장되어 있다.

3.1 DeepLApi호출을 안정적으로 하기위한 코드는 trans.py에 한꺼번에 모아두었다.

3.2 translation\_progress.pkl파일을 지우게 된 경우 번역을 처음부터 다시 시작하게 된다. 이때 사용한 DeepL의 API key는 개인 소유이기 때문에 지워져 있으므로 적절한 API key를 입력하여야 한다.

3.3 translation\_progress.pkl파일이 손상될 경우를 대비하여 updated\_data.pkl 파일에 번역 결과를 백업해 놓았다. 백업 결과를 사용하려면 options를 False로 하면 된다.

3.4 번역된 결과물은 각각 \_tr이 붙은 컬럼에 저장하였다.

4. review\_text에 ‘The media could not be loaded.\n \n\n\n\n\n \xc2\xa’을 포함하는 경우 원본 데이터에 멀티미디어 파일이 포함된 경우로 csv파일을 생성 중 생긴 오류로 추정된다. 이 구문을 찾고 지워준다.

**3. 감성 분석**

1. Stop word 제거를 하였다. 단어의 총 수가 3개 이하거나, 최종 결과물의 단어수가 3개 이하이면, 원본 데이터를 그대로 사용하도록 하였다.

2. review\_title, review\_text에 적용하였다.

**4. 모델 선택 및 학습**

1. 0,8/0.2의 비율로 테스트 데이터를 생성하였다. 최종 예측 목표는 review\_rating이다.

2. 선형분석, Ridge 등 총 9개의 모델에 대하여 분석을 실행하였다.

분석 결과의 순위는 아래와 같다.

1. Gradient Boosting

2. Random Forest

3. Linear Regression, Ridge

4. Lasso, Elastic Net

3. 위의 결과를 토대로 Gradient Boosing을 분석에 사용하기로 결정하고 베이지안 최적화를 통해 하이퍼 파라미터 튜닝을 진행하였다.

4. 최종 결과의 정확도는 아래와 같다. Test 데이터를 사용하여 성능을 측정한 결과와 Train데이터를 사용하여 성능을 측정한 경우 2가지 경우를 확인해 보았다.

Test MSE: 1.441

Test R^2: 0.218

Test RMSE: 1.200

Test MAE: 0.896

Test MAPE: 0.412

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Train MSE: 1.128

Train R^2: 0.363

Train RMSE: 1.062

Train MAE: 0.799

Train MAPE: 0.356

**4. 분석 결과**

1. RMSE값이 1.2라는 것은 예측 값이 1.2 정도의 오차를 가진다는 것이다.

예측하고자 하는 값이 평점 즉 0~5 사이의 값을 가지므로 오차가 예상 성능보다 크다

R² 값 또한 21.6%로 데이터의 분산이 큰 것 또한 알 수 있다.

2. 오버피팅

위에서 훈련 데이터와 테스트 데이터에서의 여러 수치들을 보여주고 있다.

전반적으로 훈련 테스트 데이터에서 더 낮은 지표를 보여준다.

이것은 오버피팅되었을 가능성이 있다는 것을 보여준다.

하지만 오버피팅된 정도는 큰 것으로 보이지는 않는다.

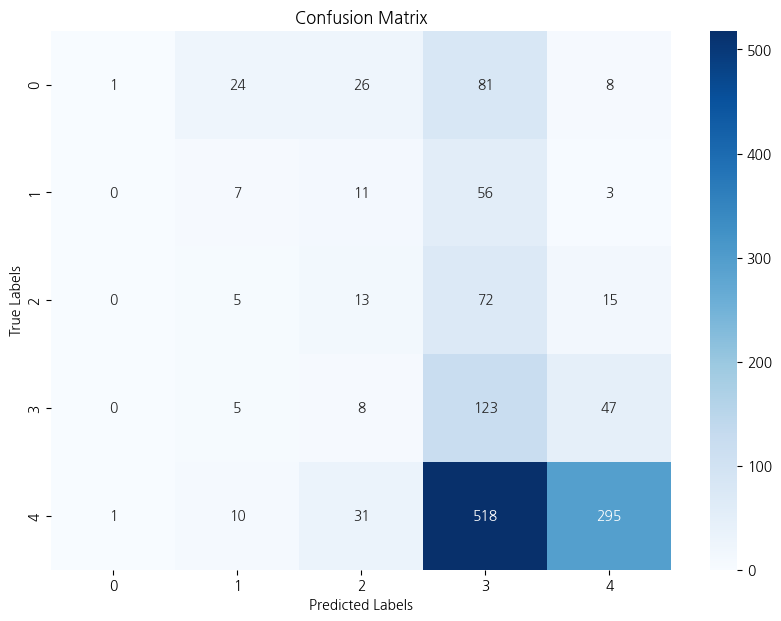
**5. 시각화**

1. 혼돈행렬

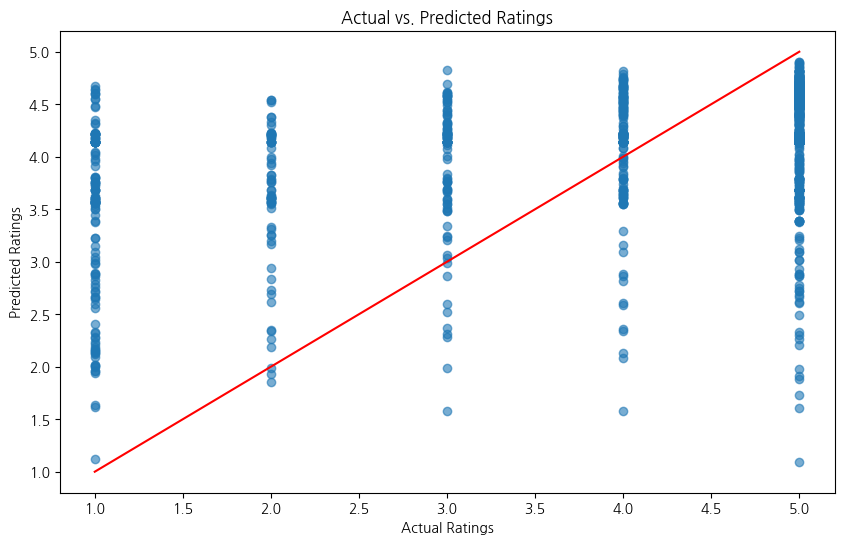
0~4점으로 평가되어 있으나 실제 데이터는 1~5점으로 각각 1점이 더해져야 하는 것에 주의하자.

1~3점 사이의 값들은 대부분 3점으로 평가되어 있다.

5점 리뷰는 5점으로 평가 경우가 295, 4점으로 평가된 경우가 518으로 그나마 높은 정확도를 보여준다.



2. Roc 그래프



모델의 예측 결과과 유의미한 수준으로 보이지 않는다. 단 5점 리뷰인 경우 그나마 정확한 것으로 보인다.