

HyperLearning_EssayEvaluation

결과 보고서

하이퍼러닝 과제전형
지원자 인재현

제출 기한
2024.12.23

데이터 분석 결과

데이터 특성

데이터 Info

Data columns (total 37 columns):			
#	Column	Non-Null Count	Dtype
0	paragraph	3085 non-null	object
1	correction	3085 non-null	object
2	score.paragraph_score	3085 non-null	object
3	score.essay_scoreT	3085 non-null	object
4	score.essay_scoreT_avg	3085 non-null	float64
5	score.essay_scoreT_detail.essay_scoreT_org	3085 non-null	object
6	score.essay_scoreT_detail.essay_scoreT_cont	3085 non-null	object
7	score.essay_scoreT_detail.essay_scoreT_exp	3085 non-null	object
8	student.date	3085 non-null	object
9	student.student_educated	3085 non-null	bool
10	student.student_grade	3085 non-null	object
11	student.location	3085 non-null	object
12	student.student_grade_group	3085 non-null	object
13	student.student_reading	3085 non-null	int64
14	rubric.essay_grade	3085 non-null	object
15	rubric.organization_weight.org_paragraph	3085 non-null	int64
16	rubric.organization_weight.org	3085 non-null	int64
17	rubric.organization_weight.org_essay	3085 non-null	int64
18	rubric.organization_weight.org_coherence	3085 non-null	int64
19	rubric.organization_weight.org_quantity	3085 non-null	int64
...
35	info.essay_len	3085 non-null	int64
36	info.essay_main_subject	3085 non-null	object

dtypes: bool(1), float64(1), int64(16), object(19)

데이터 Features

```
[('paragraph', 'correction', 'score.paragraph_score',  
'score.essay_scoreT', 'score.essay_scoreT_avg',  
'score.essay_scoreT_detail.essay_scoreT_org',  
'score.essay_scoreT_detail.essay_scoreT_cont',  
'score.essay_scoreT_detail.essay_scoreT_exp', 'student.date',  
'student.student_educated', 'student.student_grade', 'student.location',  
'student.student_grade_group', 'student.student_reading',  
'rubric.essay_grade', 'rubric.organization_weight.org_paragraph',  
'rubric.organization_weight.org',  
'rubric.organization_weight.org_essay',  
'rubric.organization_weight.org_coherence',  
'rubric.organization_weight.org_quantity', 'rubric.essay_type',  
'rubric.content_weight.con_clearance',  
'rubric.content_weight.con_novelty', 'rubric.content_weight.con',  
'rubric.content_weight.con_prompt',  
'rubric.content_weight.con_description', 'rubric.essay_main_subject',  
'rubric.expression_weight.exp_style',  
'rubric.expression_weight.exp_grammar',  
'rubric.expression_weight.exp_vocab', 'rubric.expression_weight.exp',  
'info.essay_id', 'info.essay_prompt', 'info.essay_type',  
'info.essay_level', 'info.essay_len', 'info.essay_main_subject'],
```

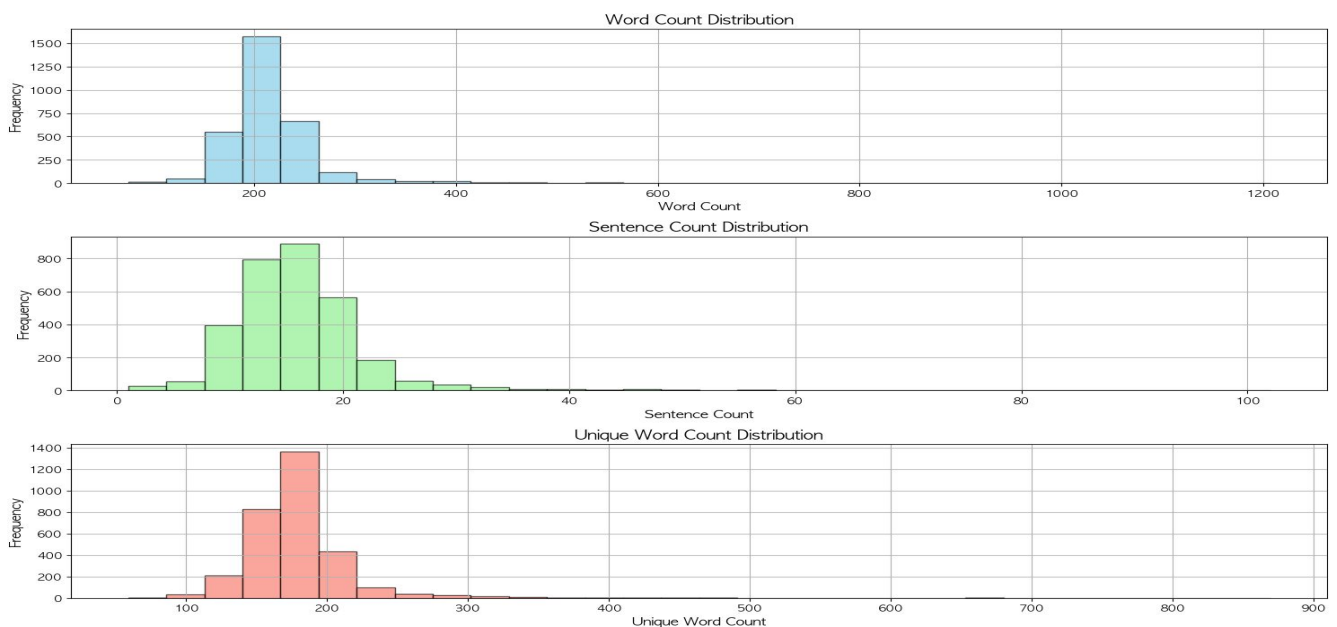
- 고등학교 2학년 essays 데이터 추출
- NaN 값 없고 우리가 필요한 Feature는 paragraph, {org, cont, exp} score 입니다.

Raw Data

	paragraph	score.essay_scoreT_detail.essay_scoreT_org	score.essay_scoreT_detail.essay_scoreT_cont	score.essay_scoreT_detail.essay_scoreT_exp
0	[('paragraph_txt': '지금까지 우주에 대해서 배운다는건 너무 어렵고 ...	[[3, 3, 3, 3], [3, 3, 3, 3], [2, 2, 2, 3]]	[[3, 3, 0, 3], [3, 3, 0, 3], [3, 2, 0, 2]]	[[2, 3, 3], [3, 3, 3], [3, 3, 3]]
1	[('paragraph_txt': '저는 우주에는 외계인이 있다고 생각합니다. #@...	[[3, 3, 3, 3], [3, 2, 3, 3], [3, 3, 3, 3]]	[[3, 3, 0, 3], [2, 3, 0, 3], [2, 3, 0, 3]]	[[3, 3, 3], [2, 2, 3], [2, 2, 3]]
2	[('paragraph_txt': '인터넷에 돌아다니다가 '우주의 소리'라는 제목의...	[[2, 3, 3, 3], [2, 2, 3, 3], [2, 2, 3, 3]]	[[3, 3, 0, 3], [3, 3, 0, 2], [2, 3, 0, 3]]	[[3, 3, 3], [3, 3, 3], [2, 2, 3]]
3	[('paragraph_txt': '우리가 살고있는 지구, 지구를 포함하고 있는 우...	[[2, 2, 2, 3], [3, 2, 2, 3], [3, 3, 3, 3]]	[[2, 2, 0, 2], [3, 2, 0, 3], [3, 2, 0, 2]]	[[2, 2, 2], [2, 3, 2], [1, 3, 3]]
4	[('paragraph_txt': '화성은 오래전부터 인류가 관심을 가져온 행성이었...	[[3, 3, 3, 3], [3, 3, 2, 3], [3, 3, 3, 3]]	[[3, 3, 0, 3], [3, 3, 0, 2], [3, 3, 0, 3]]	[[3, 3, 3], [2, 3, 3], [3, 3, 3]]
...
3080	[('paragraph_txt': '컴퓨터와 정보 통신 기술이 발달이 되어 편리한 ...	[[3, 3, 3, 3], [3, 3, 3, 3], [3, 3, 3, 3]]	[[3, 3, 3, 3], [3, 3, 3, 3], [3, 3, 3, 3]]	[[3, 3, 3], [3, 3, 3], [3, 3, 3]]
3081	[('paragraph_txt': '제한적 인터넷 실명제 필요성에 대하여 반대하는 ...	[[3, 3, 3, 3], [3, 2, 3, 3], [3, 3, 3, 3]]	[[3, 2, 3, 3], [3, 3, 3, 2], [3, 3, 3, 3]]	[[3, 3, 3], [3, 3, 3], [3, 3, 3]]
3082	[('paragraph_txt': '인터넷 실명제는 실행되지 않아야 한다. #@문장...	[[3, 3, 3, 3], [3, 3, 3, 3], [3, 3, 3, 3]]	[[3, 3, 3, 3], [2, 3, 3, 3], [3, 3, 3, 3]]	[[3, 3, 3], [3, 3, 3], [3, 3, 3]]
3083	[('paragraph_txt': '인터넷 실명제에 반대하는 입장이다. #@문장구분#...	[[3, 3, 3, 3], [3, 3, 3, 3], [3, 3, 3, 3]]	[[3, 2, 3, 3], [3, 2, 3, 3], [3, 3, 3, 3]]	[[3, 3, 3], [3, 3, 3], [3, 3, 3]]
3084	[('paragraph_txt': '익명성에 숨어 악성댓글 등으로 특정 대상을 집단...	[[3, 3, 3, 3], [2, 2, 2, 3], [1, 1, 1, 1]]	[[2, 2, 2, 3], [2, 2, 3, 2], [1, 1, 1, 1]]	[[2, 3, 3], [2, 2, 2], [1, 1, 1]]

3085 rows x 4 columns

텍스트 데이터 길이 분포 (단어 수, 문장 수, 고유 단어 수)



→ 고등학교 2학년 에세이는 단어 수가 200단어 전후로 분포가 높음을 관찰할 수 있습니다.

데이터 전처리 결과(불용어 제거)

Paragraph

```
0      지금까지 우주에 대해서 배운다는건 너무 어렵고 이해하기 쉽지 않아서 힘들었는데 해성...
1      저는 우주에는 외계인이 있다고 생각합니다. 그 이유는 우주는 우리 생각보다 더 크...
2      인터넷에 돌아다니다가 '우주의 소리'라는 제목의 영상을 접하게 된 적이 있는데 목성...
3      우리가 살고있는 지구, 지구를 포함하고 있는 우주, 이런것들이 생겼기에 인체의신비,...
4      화성은 오래전부터 인류가 관심을 가져온 행성이었습니다. 인류가 달에 착륙한 지 오...
...
3080   컴퓨터와 정보 통신 기술이 발달이 되어 편리한 의사 소통과 다양한 자료의 확보, 물...
3081   제한적 인터넷 실명제 필요성에 대하여 반대하는 입장입니다. 인터넷 실명제라는 것은...
3082   인터넷 실명제는 실행되지 않아야 한다. 인터넷 실명제의 문제점은 자유가 침해된다....
3083   인터넷 실명제에 반대하는 입장이다. 첫째 인터넷 실명제가 표현의 자유를 억압할 수...
3084   익명성에 숨어 악성댓글 등으로 특정 대상을 집단적으로 따돌리거나 집요하게 괴롭히는'...
Name: cleaned_paragraph, Length: 3085, dtype: object
```

구성 점수

내용 점수

표현 점수

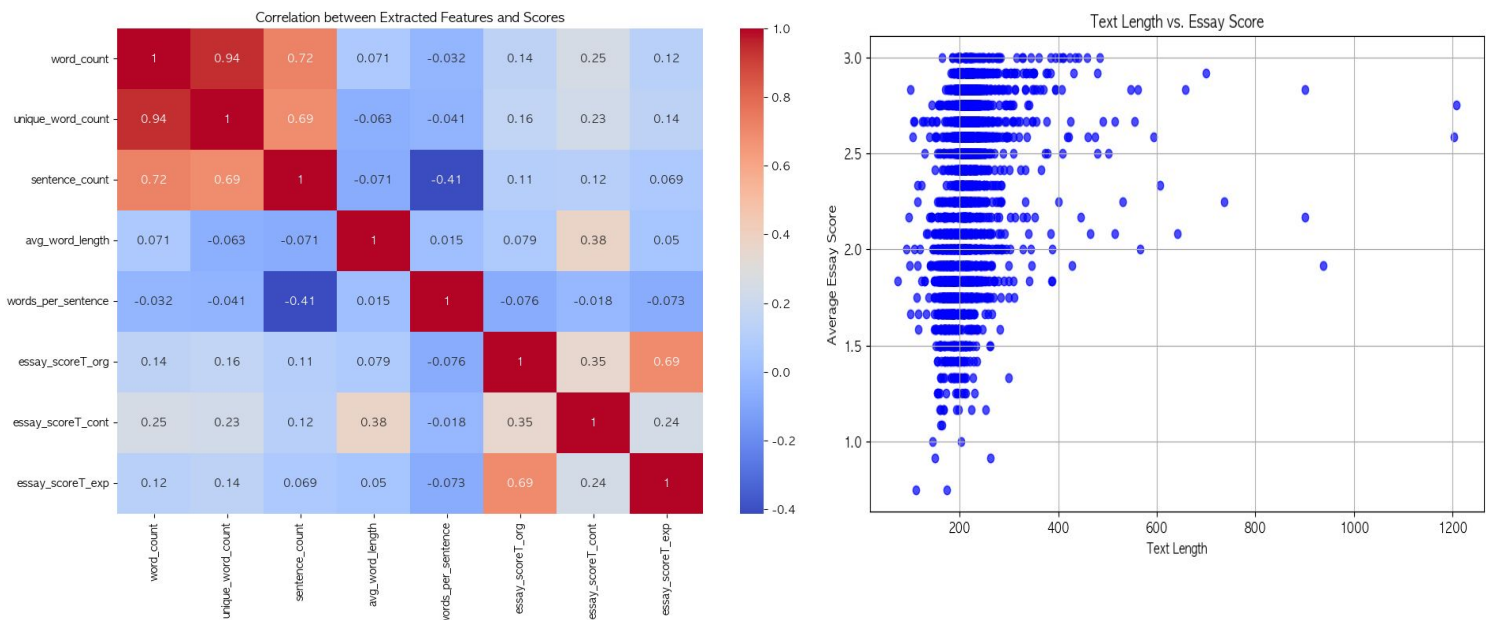
essay_scoreT_org	essay_scoreT_cont	essay_scoreT_exp
2.750000	2.083333	2.888889
2.916667	2.083333	2.555556
2.583333	2.083333	2.777778
2.583333	1.750000	2.222222
2.916667	2.166667	2.888889

→ 리스트 형태의 데이터는 분석 및 모델링에 직접적으로 사용이 어려워 **단일 스칼라 형태로 변환**해야합니다. 각 리스트의 **평균 값**을 대표적인 값으로 나타냈습니다.

정규화 및 토큰화

cleaned_paragraph	normalized_paragraph	tokenized_paragraph
지금까지 우주에 대해서 배운다는건 너무 어렵고 이해하기 쉽지 않아서 힘들었는데 해성...	지금까지 우주에 대해서 배운다는건 너무 어렵고 이해하기 쉽지 않아서 힘들었는데 해성...	[_지금까지, _우주, 에, _대해서, _배, 운, 다, 는, 건, _너무, _어렵...
저는 우주에는 외계인이 있다고 생각합니다. 그 이유는 우주는 우리 생각보다 더 크...	저는 우주에는 외계인이 있다고 생각합니다 그 이유는 우주는 우리 생각보다 더 크고 ...	[_저, 는, _우주, 에는, _외, 계, 인, 이, _있다고, _생각, 합니다, ...
인터넷에 돌아다니다가 '우주의 소리'라는 제목의 영상을 접하게 된 적이 있는데 목성...	인터넷에 돌아다니다가 우주의 소리라는 제목의 영상을 접하게 된 적이 있는데 목성이나...	[_인터넷, 에, _돌아, 다, 니다, 가, _우주, 의, _소리, 라는, _제, ...
우리가 살고있는 지구, 지구를 포함하고 있는 우주, 이런것들이 생겼기에 인체의신비,...	우리가 살고있는 지구 지구를 포함하고 있는 우주 이런것들이 생겼기에 인체의신비 생명...	[_우리가, _살, 고, 있는, _지구, _지구, 를, _포함, 하고, _있는, ...
화성은 오래전부터 인류가 관심을 가져온 행성이었습니다. 인류가 달에 착륙한 지 오...	화성은 오래전부터 인류가 관심을 가져온 행성이었습니다 인류가 달에 착륙한 지 오랜 ...	[_화성, 은, _오래, 전, 부터, _인, 류, 가, _관심을, _가져, 온, ...

Feature Engineering 및 상관관계



→ 특성이 너무 적어 새로운 특성을 만들고 상관관계를 구했습니다.

“평균 단어 길이”와 “문장당 평균 단어 수”를 특성으로 만들었습니다.

“평균 단어 길이” 특성은 글의 어휘적 수준을 나타낼 수 있다 생각하였고,

“문장당 단어 수”는 글의 논리적 구성이나 가독성을 반영할 수 있다 생각하여 생성하였습니다.

→ 데이터가 **비선형적**으로 보입니다.

모델 학습

- 목표 : '내용', '구성', '표현' 점수 예측 모델 개발

첫번째 시도 - XGBoost

Fitting 5 folds for each of 54 candidates, totalling 270 fits

Best Parameters (XGBoost): {'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 200, 'subsample': 0.8}

Best CV R² Score: 0.08871254422699268

XGBoost Test R² Score: 0.0978

- **모델을 선택한 이유**

- XGBoost는 비선형적이고 복잡한 데이터를 처리하는데 강점이 있기 때문입니다.
다만 여기서 데이터가 작거나 특성이 부족해서 기대한 성능을 보이지 못했습니다.

- **성능 평가 결과**

- 1) 하이퍼 파라미터 튜닝 및 교차검증을 진행하였습니다.
성능 점수를 R² 으로 하였는데, 0에 가까울수록 모델이 데이터를 설명하지 못한다는 뜻입니다.
- 2) 0.0978의 점수로 데이터 특성이나 모델의 적합성을 재검토할 필요가 있어 보입니다.

- **결과 개선 방안**

- 1) **데이터 품질 개선**
 - a) 더 많은 데이터 확보 → 조건이 고등학교 2학년 데이터만 사용 - 현재 불가.
- 2) **다른 모델 선택**
 - a) LightGBM, RandomForest, 딥러닝 모델 시도
- 3) **Feature Engineering**
 - a) 추가적인 텍스트 기반 특성을 생성

두번째 시도 - 딥러닝 모델 - Dense Neural Network(DNN)

Mean Absolute Error (MAE): 3.8405

구성 점수	내용 점수	표현 점수
[-0.57558006	-0.19949198	0.1803594]
[-4.445265	-3.6803017	-3.090599]
[2.9009957	3.00737	3.1679401]
[4.3476996	4.270389	4.3751154]]

- **모델을 선택한 이유**

- 데이터가 정량적 특성(word_count, sentence_count 등)과 정규화된 수치 값을 포함하고 있어,
연속적인 출력값을 예측 할 수 있는 회귀 모델로 딥러닝이 적합하다 생각했습니다.
그래서 딥러닝의 다층 퍼셉트론(MLP) 구조를 활용했습니다.

- **성능 평가 결과**

- 예측된 출력값의 일부가 음수로 나타났습니다. 이는 데이터 정규화 과정에서 범위를 제한하지 않았기 때문이라고 생각합니다. 또한 현재 모델이 128-64 노드로 구성된 단순한 신경망 구조입니다.

- **결과 개선 방안**

- 1) **데이터 정규화 및 범위 제한**
 - a) 추가적인 스케일링 기법을 적용
 - b) 예측값을 실제 점수의 범위로 제한하는 함수 추가
- 2) **모델 복잡성 조정**
 - a) 모델의 레이어를 추가하거나, 노드 수를 증가시켜 더 복잡한 데이터 패턴을 학습할 수 있도록 구조를 변경.

세번째 시도 - 딥러닝 모델 - Dense Neural Network(DNN) 조정

Mean Absolute Error (MAE): 0.1399

구성 점수 내용 점수 표현 점수

Sample Predictions:

```
[[2.6294913 2.2607312 2.5900745]
 [2.60671   2.294247  2.5630584]
 [2.5748887 2.2371197 2.518658 ]
 [2.7108395 2.3861978 2.6591735]
 [2.694746  2.3745656 2.657702 ]]
```

● 모델을 선택한 이유

- 출력값의 정규화 - 0~1 범위로 정규화하고, 학습 이후 다시 원래 점수 범위로 변환하였습니다.
- 활성화 함수 변경 - sigmoid 활성화 함수를 사용하여 출력값을 0~1 범위로 제한하였습니다.

● 성능 평가 결과

- 1) 출력값이 실제 점수 범위와 가까워졌으며, MAE 값이 안정적으로 낮아졌습니다.
→ Sigmoid 활성화 함수가 출력값을 0~1로 제한하여 모델 학습이 안정적으로 이루어졌으며, EarlyStopping을 통해 과적합을 방지할 수 있었습니다.
- 2) 예측값의 분포가 타겟 점수와 유사한 패턴을 보였으며, 학습 및 평가 시 과적합 방지가 확인되었습니다.

● 한계점

- 1) 데이터 부족
→ 현재 데이터는 고등학교 2학년 에세이에 국한되어 있어 일반화 성능이 부족할 수 있습니다.
- 2) 특성의 단순성
→ 현재 사용된 특성 (단어 수, 문장 수 등)은 점수 예측에 영향을 미칠 수 있으나, 추가적인 문법적, 의미적 특성 분석이 부족합니다.

● 결과 개선 방안

- 1) 데이터 확대
 - a) 다른 학년의 데이터를 추가하여 모델의 일반화 성능을 높일 수 있습니다.
- 2) 다른 딥러닝 모델로 변경
 - a) 예를 들어 Transformer 기반 모델로 변경하거나, 하이퍼파라미터 튜닝을 통해 성능을 더욱 개선할 수 있습니다.

결론

- DNN 모델은 현재 데이터셋과 문제 정의에 적합한 성능을 보여주었으며, 최종적으로 0.1399의 MAE를 기록했습니다.
- 현재 데이터와 특성으로는 점수 예측에서 일정 수준의 성과를 냈지만, 추가적인 데이터와 특성 확장이 필요합니다.
- 향후 연구로는 Transformer 기반 모델로의 변경 및 하이퍼파라미터 튜닝을 통한 성능 개선을 제안합니다.