모델 성능 분석 보고서

다양한 회귀 모델을 사용하여 에세이 점수를 예측하고자 함.

사용한 모델: **Linear Regression**, **Random Forest Regressor**, **Support Vector Regressor**, **ElasticNet**.

각 모델의 성능을 MSE와 교차 검증 결과를 통해 평가하고, 각 모델의 한계점과 개선 가능성을 분석함.

데이터셋 정보

1. info: 에세이 기본 정보
    1.1. essay_id: 에세이 아이디
    1.2. essay_type: 에세이 유형
    1.3. essay_main_subject: 에세이 주제
    1.4. essay_level: 에세이 난이도
    1.5. essay_prompt: 에세이 프롬프트
    1.6. essay_len: 에세이 글자수

2. student: 에세이 작성자 정보
    2.1. student_grade_group: 학생 학년군
    2.2. student_grade: 학생 학년
    2.3. student_reading: 학생이 일주일 간 읽은 책의 양
    2.4. student_educated: 학생의 논술 사교육 유무
    2.5. date: 에세이 수집일
    2.6. location: 에세이 수집 장소

3. paragraph: 에세이 문단별 정보 목록
    3.1. paragraph_id: 문단 아이디
    3.2. paragraph_txt: 문단 텍스트
    3.3. paragraph_len: 문단 글자수

4. score: 에세이 점수 정보
    4.1. essay_scoreT: 3명의 평가자 에세이 총점 목록
    4.2. essay_scoreT_avg: 3명의 평가자 에세이 총점 평균
    4.3. essay_scoreT_detail: 항목별 세부 점수
        4.3.1. essay_scoreT_exp: 표현 점수 목록
        4.3.2. essay_scoreT_org: 구성 점수 목록
        4.3.3. essay_scoreT_cont: 내용 점수 목록
    4.4. paragraph_score: 문단별 점수 정보

5. rubric: 루브릭 정보
    5.1. rubric_essay_type: 에세이 유형
    5.2. rubric_essay_main_subject: 에세이 주제
    5.3. rubric_essay_grade: 에세이 작성자 학년

6. expression_weight: 루브릭 표현 가중치 정보

    6.1. exp: 대분류 표현 가중치
    6.2. exp_grammar: 문법 가중치
    6.3. exp_vocab: 단어 가중치
    6.4. exp_style: 문장 표현 가중치

7. organization_weight: 루브릭 구성 가중치 정보
    7.1. org: 대분류 구성 가중치
    7.2. org_essay: 문단 간 가중치
    7.3. org_paragraph: 문단 내 가중치
    7.4. org_coherence: 일관성 가중치
    7.5. org_quantity: 분량 가중치

8. content_weight: 루브릭 내용 가중치 정보
    8.1. con: 대분류 내용 가중치
    8.2. con_clearance: 주제 명료성 가중치
    8.3. con_novelty: 참신성 가중치
    8.4. con_prompt: 프롬프트 독해력 가중치
    8.5. con_description: 서술력 가중치

---

1. 선택한 모델과 그 이유

선택한 모델

- **Linear Regression**
- **Random Forest Regressor**
- **Support Vector Regressor**
- **ElasticNet**

모델 선택 이유

- **Linear Regression**: 가장 기본적인 회귀 모델로, 다른 모델들과 비교할 때 모델 해석이 용이하고 훈련 속도가 빠름. 데이터셋에서 변수 간의 관계가 선형적일 가능성이 있어 이를 먼저 시도함.
- **Random Forest Regressor**: 비선형 관계를 모델링할 수 있는 앙상블 기법으로, 여러 결정 트리를 결합하여 예측 성능을 향상시킬 수 있음. 특히 고차원 데이터에서 좋은 성능을 보이며, feature importance 분석을 통해 중요한 변수를 파악할 수 있음.
- **Support Vector Regressor**: 비선형 데이터에 적합한 모델로, 커널 함수를 사용해 복잡한 데이터 구조를 잘 처리할 수 있음. 작은 데이터셋에서 효과적이고 고차원 데이터를 다룰 때도 성능이 우수함.
- **ElasticNet**: L1(라쏘)와 L2(릿지) 정규화를 결합한 모델로, 데이터 내에서 상관성이 있는 변수들을 잘 처리할 수 있음. 특히 과적합을 방지하는 데 유리한 모델이기 때문에 선택함.

---

2. 성능 평가 결과

모델	Test Set MSE	교차 검증 평균 MSE	교차 검증 표준편차
Linear Regression	5.76	5.77e+26	8.22e+26
Random Forest Regressor	6.01	6.99	1.36
Support Vector Regressor	5.95	6.06	0.61
ElasticNet	6.53	5.93	0.96

분석 및 성능 해석

- **Linear Regression**: MSE 값이 상당히 높고, 교차 검증에서 나온 값들이 매우 커서 모델이 잘못 학습되었을 가능성이 큼. 특히 교차 검증 MSE 값이 너무 커서 과적합(overfitting)이나 학습 실패가 의심됨.
  - **MSE 값이 큰 이유**: 선형 회귀 모델이 데이터의 비선형적 관계를 잘 반영하지 못했거나, 과적합이 발생했을 수 있음. 또한, 데이터 스케일링 문제나 특성 선택이 부적절했을 가능성도 있음.
- **Random Forest Regressor**: Test Set에서 MSE가 6.01로 비교적 낮고, 교차 검증에서의 평균 MSE는 6.99로 안정적인 성능을 보임. 표준편차도 낮아 모델이 예측에서 일관성을 유지한 것으로 판단됨.
- **Support Vector Regressor (SVR)**: Test Set MSE가 5.95로 낮고, 교차 검증에서의 평균 MSE가 6.06으로 안정적인 성능을 나타냄. 표준편차도 0.61로 낮아 예측의 일관성이 높다고 볼 수 있음.
- **ElasticNet**: Test Set MSE는 6.53으로 다른 모델에 비해 다소 높은 값이지만, 교차 검증에서는 5.93으로 평균 MSE가 낮고 표준편차도 안정적임. ElasticNet은 과적합을 방지하면서도 좋은 성능을 보인 모델임.

---

3. 한계점 및 개선 방안

한계점

- **선형 회귀 모델의 성능**: Linear Regression은 데이터의 비선형 관계를 잘 반영하지 못했을 가능성이 높음. 또한, 데이터의 스케일이나 특성 선택이 모델 성능에 부정적인 영향을 미쳤을 수 있음.
- **모델 과적합**: 특히 Linear Regression에서 비정상적으로 큰 MSE 값이 나온 것은 과적합이 발생했을 가능성 큼. 모델이 훈련 데이터에 너무 치우쳐 학습하면서 새로운 데이터에 대해 잘 일반화되지 않았을 수 있음.
- **특성 선택 문제**: 중요한 특성들이 누락되었을 수 있음. 추가적인 feature engineering을 통해 더 많은 특성을 고려하거나, 중요하지 않은 특성들을 제거하여 성능을 개선할 수 있음.

개선 방안

1. **데이터 전처리 개선**
   - **스케일링 기법 개선**: 현재는 MinMaxScaler를 사용하고 있지만, StandardScaler로 실험을 진행할 수 있음.
   - **Feature Engineering**: 기존 특성 외에 문단 길이와 에세이 길이를 결합한 새로운 특성을 추가하거나, 다른 변형된 특성을 생성할 수 있음.
2. **모델 튜닝**
   - **하이퍼파라미터 튜닝**: Random Forest와 SVR 성능 개선을 위해 GridSearchCV나 RandomizedSearchCV로 하이퍼파라미터를 최적화할 수 있음.
   - **고급 모델 실험**: XGBoost, LightGBM과 같은 부스팅 모델을 사용하여 성능을 향상시킬 수 있음. 이들 모델은 랜덤 포레스트보다 더 좋은 예측 성능을 보일 수 있음.
3. **교차 검증 추가**
   - 교차 검증을 더 많이 진행하면 모델 성능을 더 신뢰성 있게 평가할 수 있음. 현재는 cv=5로 설정했지만, cv=10으로 확장해 평가를 강화할 수 있음.

결론

- **최적 모델**: **Random Forest Regressor**와 **Support Vector Regressor**가 우수한 성능을 보였으며, 특히 **Support Vector Regressor**가 예측 성능과 안정성 모두에서 더 뛰어난 결과를 보임.
- **성능 향상 가능성**: 선형 회귀는 비선형 관계를 잘 반영하지 못해, 비선형 모델인 **Random Forest**나 **SVR**이 더 적합한 선택이었음.
- **향후 개선 사항**: 모델 튜닝, 데이터 전처리 개선, 새로운 모델 실험을 통해 성능을 더욱 향상시킬 수 있음.