📊 Kaggle Playground S6E1 - 학생 시험 점수 예측

EDA 인사이트 정리 및 모델링 방향성

---

1. 데이터 개요

항목	내용
데이터 크기	630,000개 샘플
Target	exam_score (회귀 문제)
결측치	없음 ✓
Target 분포	평균 62.5점, 표준편차 18.9점 (19.6 ~ 100점)

---

2. 핵심 인사이트 (변수 중요도 순)

🥇 1순위: study_hours (공부 시간)

- **상관계수: 0.7623** (매우 강한 양의 상관관계)
- 시험 점수의 약 58%를 설명하는 가장 중요한 변수
- **결론: 모델의 핵심 피처, 반드시 포함**

🥈 2순위: sleep_quality (수면의 질)

- **Kruskal H: 33,168** (범주형 중 가장 강한 영향력)
- Poor → Average → Good → Excellent 순으로 점수 상승
- 흥미로운 발견: **수면 시간(0.17)보다 수면의 질이 훨씬 중요**
- 수면 시간과 수면의 질 상관계수: 0.0255 (거의 무관!)

🥉 3순위: study_method (학습 방법)

- **Kruskal H: 29,548**
- 5가지 학습 방법 간 점수 차이가 매우 큼
- coaching, group study, online videos, self-study, tutoring 중 어떤 것이 효과적인지 추가 분석 필요

4순위: facility_rating (시설 평가)

- **Kruskal H: 20,684**
- Low → Medium → High 순으로 점수 상승
- 학습 환경이 성적에 유의미한 영향

5순위: class_attendance (출석률)

- **상관계수: 0.3610** (중간 수준의 상관관계)
- 출석을 잘 할수록 성적이 좋음

6순위: sleep_hours (수면 시간)

- **상관계수: 0.1674** (약한 상관관계)
- 수면의 질보다 영향력이 낮음

⚠️ **영향력 낮은 변수들**

- **age**: 상관계수 0.0105 (거의 무관, R^2 = 0.0001)
- **internet_access**: ANOVA p-value 0.7226 (유의미하지 않음!)
- **exam_difficulty**: H=53.9536 (상대적으로 낮은 영향)
- **gender, course**: 유의미하나 영향력 상대적으로 낮음

---

3. 핵심 발견 & 비즈니스 인사이트

💡 **발견 1: "양보다 질"**

> 수면은 **시간보다 질**이 중요하다!
> - 수면 시간 → 시험 점수: 상관계수 0.17 (약한 영향)
> - 수면의 질 → 시험 점수: Kruskal H 33,168 (매우 강한 영향)
> - 수면 시간 ↔ 수면의 질: 상관계수 0.0255 (거의 무관)

💡 **발견 2: 인터넷 접근성은 무관**

- internet_access가 시험 점수에 유의미한 영향을 미치지 않음
- 모델링 시 제거 또는 낮은 가중치 고려

💡 **발견 3: 나이는 성적과 거의 무관**

- 17~24세 범위에서 나이에 따른 성적 차이가 거의 없음
- Feature Engineering 시 제거 가능

4. 모델링 방향성 제안

📌 Phase 1: Baseline 모델

추천 모델: LightGBM, XGBoost, CatBoost
평가 지표: R^2 또는 RMSE (Playground Series 표준)

📌 Phase 2: Feature Engineering

A. 생성할 피처들

새 피처	설명	근거
study_efficiency	study_hours / sleep_hours	수면 대비 공부 효율
study_attendance_ratio	study_hours × class_attendance	공부 + 출석 시너지
quality_hours	sleep_quality_encoded × sleep_hours	수면 질×양 상호작용
total_study_quality	study_hours × facility_rating_encoded	공부시간 × 환경

B. 인코딩 전략

```python
# 순서형 변수 (Ordinal Encoding)
sleep_quality: poor=1, average=2, good=3, excellent=4
facility_rating: low=1, medium=2, high=3
exam_difficulty: easy=1, moderate=2, hard=3

# 명목형 변수 (Target Encoding 또는 One-Hot)
gender, course, study_method
```

C. 제거/가중치 하향 고려 피처

- internet_access (유의미하지 않음)
- age (거의 영향 없음)

📌 Phase 3: 앙상블 전략

1단계: 3가지 GBDT 모델 (LGBM, XGB, CatBoost) 개별 학습
2단계: Weighted Average 또는 Stacking
3단계: Optuna로 하이퍼파라미터 튜닝

📌 Phase 4: 교차 검증 전략

```python
# 추천: K-Fold Cross Validation
from sklearn.model_selection import KFold

kf = KFold(n_splits=5, shuffle=True, random_state=42)
# 63만개 데이터 → 5-Fold 적당
```

---

5. 추가 분석 제안

✅ 아직 분석하지 않은 것들

1. **study_method별 상세 분석**
   - 어떤 학습 방법이 가장 효과적인지?
   - study_hours와 study_method의 상호작용 효과
2. **다중공선성 체크**
   - VIF (Variance Inflation Factor) 확인
   - 특히 sleep_hours와 다른 변수들 간 관계
3. **이상치 분석**
   - exam_score 19.6점 (최저)인 학생들의 특성
   - study_hours 7.91시간 (최대)인데 점수 낮은 케이스
4. **비선형 관계 탐색**
   - study_hours와 exam_score의 비선형 패턴
   - 다항 피처 고려
5. **그룹별 세분화 분석**
   - course별로 다른 요인이 중요할 수 있음
   - 전공별 모델 또는 전공 피처 상호작용

---

6. 실험 우선순위 체크리스트

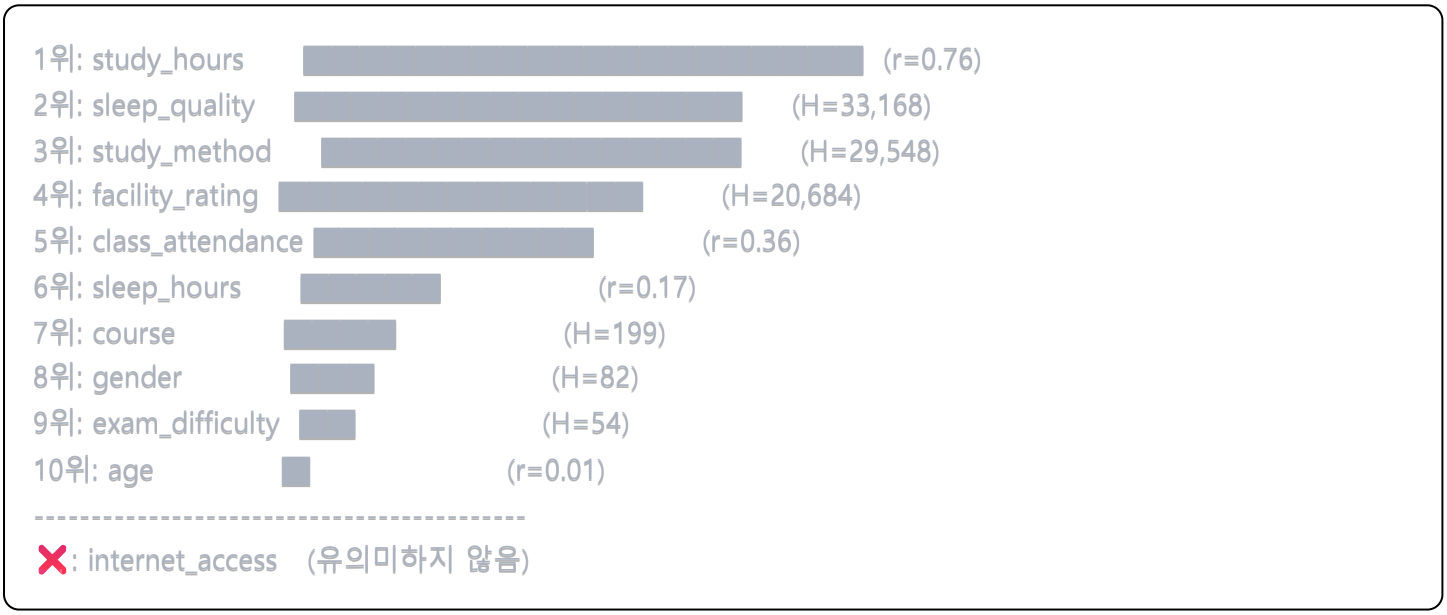
우선순위	작업	예상 효과
⭐⭐⭐	LightGBM baseline + 기본 피처	빠른 baseline 확보
⭐⭐⭐	study_hours × sleep_quality 상호작용	핵심 피처 시너지

우선순위	작업	예상 효과
⭐⭐⭐	순서형 변수 Ordinal Encoding	모델 성능 향상
⭐⭐	study_method Target Encoding	범주형 처리 개선
⭐⭐	LGBM + XGB + CatBoost 앙상블	성능 안정화
⭐	internet_access, age 제거 실험	노이즈 제거
⭐	2차 다항 피처 추가	비선형 캡처

---

7. 요약

📊 변수 중요도 랭킹 (최종)



🎯 핵심 전략

1. **공부 시간(study_hours)** 중심의 모델 구축
2. **수면의 질(sleep_quality)** 활용 (수면 시간보다 중요!)
3. **학습 방법(study_method)** + **시설 평가(facility_rating)** 상호작용
4. **internet_access, age는 노이즈**로 간주하고 실험

분석 기준일: 2026.01.20 데이터: Kaggle Playground Series S6E1 Train Set (630,000 samples)