2024-12-16 수상작 리뷰

[머신러닝 기초 트랙 시즌5] 학생 성적 예측 AI 해커톤

https://dacon.io/competitions/official/236378/codeshare/12112

데이터

1. train.csv : 학습 데이터

- id: 데이터 고유 id
- gender : 성별 (M = 남자, F = 여자)
- age: 나이
- address : 거주 지역
  U : 도시
  R : 외곽
- famsize : 가족 규모
  LE3 : 3명 이하
  GT3 : 3명 초과
- Pstatus : 부모와 동거 여부
  T : 함께 사는 경우
  A : 떨어져 사는 경우
- M_education : 어머니의 최종학력
  1 : 석박사
  2 : 초대졸 및 대졸
  3 : 고졸
  4 : 고졸 이하
- F_education : 아버지의 최종학력
  1 : 석박사
  2 : 초대졸 및 대졸
  3 : 고졸
  4 : 고졸 이하
- Mjob : 어머니의 직업
  at_home : 주부
  services : 서비스업
  teacher : 교육업
  health : 의료업
  other : 기타
- Fjob : 아버지의 직업
  at_home : 주부
  services : 서비스업
  teacher : 교육업

health : 의료업
other : 기타

- relationship_breakdown : 가족 관계
  Yes : 좋지않음
  No : 좋음
- tution_fee : 수업료
- avg_friend_hours : 하루 평균 친구와 보내는 시간
- fs_result : 1학기 성적
- avg_sleep_hours : 하루 평균 수면 시간
- avg_smartphone_hours : 하루 평균 스마트폰 사용 시간
- ss_result : 2학기 성적 (target)

2. test.csv : 테스트 데이터

- id: 데이터 고유 id
- gender : 성별 (M = 남자, F = 여자)
- age: 나이
- address : 거주 지역
  U : 도시
  R : 외곽
- famsize : 가족 규모
  LE3 : 3명 이하
  GT3 : 3명 초과
- Pstatus : 부모와 동거 여부
  T : 함께 사는 경우
  A : 떨어져 사는 경우
- M_education : 어머니의 최종학력
  1 : 석박사
  2 : 초대졸 및 대졸
  3 : 고졸
  4 : 고졸 이하
- F_education : 아버지의 최종학력
  1 : 석박사
  2 : 초대졸 및 대졸
  3 : 고졸
  4 : 고졸 이하
- Mjob : 어머니의 직업
  at_home : 주부
  services : 서비스업
  teacher : 교육업

health : 의료업
other : 기타

- Fjob : 아버지의 직업
  at_home : 주부
  services : 서비스업
  teacher : 교육업
  health : 의료업
  other : 기타
- relationship_breakdown : 가족 관계
  Yes : 좋지않음
  No : 좋음
- tution_fee : 수업료
- avg_friend_hours : 하루 평균 친구와 보내는 시간
- fs_result : 1학기 성적
- avg_sleep_hours : 하루 평균 수면 시간
- avg_smartphone_hours : 하루 평균 스마트폰 사용 시간

3. sample_submission.csv : 제출 양식

- id: 데이터 고유 id
- ss_result : 2학기 성적(예측값)

코드 흐름

1) EDA

패키지 설치
info(), describe(), head()를 통해 데이터 살펴보기
msno.matrix(train) 확인
범주형 변수의 분포를 시각화해서 살펴보기
수치형 변수의 히스토그램 살펴보기
히트맵 살펴보기

2) Feature Engineering & Processing

결측치 채우기

- 나이는 20~24 사이로, 성적은 예측해서 채워준다.
- 어머니가 가정주부인 가정이 많기 때문에 부모님의 직업을 at_home을 0, 나머지를 2로 수치형 변환한다.
- 수업료는 대체로 100만원 사이에 분포해 있으니 100고 100이상 정도로 분류한다.
- 친구와 보내는 시간은 1시간이 가장 많으므로 1시간 이하와 1시간 이상으로 분류한다.
- 1학기 성적, 스마트폰으로 비슷한 방식으로 분류한다.

타겟 및 피처 정의

원핫 인코딩 및 스케일링
RandomForest로 feature selection 적용
fs_result 모델링
전체 train 데이터 통합
다시 원핫 인코딩 / 스케일링
결측치 예측
스케일링

3) **모델 학습(Linear Regression)**

RandomForestRegressor, XGBRegressor, LGBMRegressor, CatBoostRegressor, 릿지로 스태깅 모델을 만든다. 최종 모델은 릿지 모델이다.
전체 train 데이터에 대한 예측
테스트 데이터 예측

4) **후처리**

예측값 할당

차별점, 배울점

데이터의 특성을 잘 파악하기 위해 노력했다. 각 데이터의 분포와 특징을 다양한 방법으로 관찰하였고, 그리고 적절히 전처리 하였다. CatBoostRegressor를 사용하여 스태킹 모델을 만들었다. CatBoostRegressor는 Gradient Boosting 기반의 머신러닝에서 제공되는 회귀 모델로, 범주형 변수를 처리하는 것에 강점이 있다.