AI 기계학습을 이용한
실내 거주자 재실 인원 탐지

아이디어 요약본

정민준
32184074 단국대학교 컴퓨터공학과
2023년 5월 31일

1. **데이터 분석(EDA):**
    A. 재실인원이 0인 데이터와 0이 아닌 데이터(1,2,3,4,5,6)로 구분하여 시각화하여 결정 경계가 명확한 것을 확인했고, 재실인원과 다른 특성들 간의 상관관계를 확인했다.
    B. 0을 제외한 재실인원이 1,2,3,4,5,6인 데이터에 대해서 시각화하여 결정 경계가 명확하지 않은 것을 확인했고, 재실인원과 다른 특성들 간의 상관관계를 확인했다.

2. **데이터 전처리**
    A. 학습 데이터의 결측치는 주로 재실인원 값과 regdate 값만 존재하므로 결측치를 전부 삭제했다.
    B. 평가 데이터의 결측치는 학습 데이터의 해당 특성의 평균값으로 대체했다.
    C. 학습 데이터의 클래스 불균형을 해소하기 위해 ADASYN을 사용하여 오버샘플링을 적용했다.

3. **특성 선택(Feature Selection)**
    A. “Pandas Profiling"을 통해 얻은 특성 간의 상관관계를 확인했다. 재실인원과 상관관계가 높은 특성들 중 다중공선성을 고려하여 특성을 선택했다.
    B. “재실인원이 0인 데이터와 0이 아닌 데이터(1,2,3,4,5,6)"와 "0을 제외한 재실 인원이 1,2,3,4,5,6인 데이터"의 경우 서로 다른 특징을 갖기 때문에 선택되는 특성도 다르다.

4. **모델 선택**
    A. 정형 데이터 분류 문제로 접근하여 Catboost Classifier 모델을 선택했다.

5. **모델 제작**
    A. EDA를 기반으로 재실인원을 “0과 0이 아닌 것으로 이진 분류”하는 모델을 제작했다.
    B. EDA를 기반으로 “이진 분류 모델이 0이 아닌 것으로 분류한 데이터”에 대해 “1, 2, 3, 4, 5, 6 중 하나로 재분류"하는 다중 분류 모델을 제작했다.
    C. 테스트 데이터는 두 모델을 모두 거쳐 최종 결과를 도출한다.

6. **Optuna를 통한 하이퍼파라미터 튜닝**
    A. 두 모델에 대해 Optuna를 활용하여 최적의 하이퍼파라미터를 탐색했다.