241108 수상작리뷰

[구내식당 식수 인원 예측 AI경진대회]

구내식당 식수 인원 예측 AI 경진대회 - DACON

❖ 데이터

a. **train.csv:**
일자, 요일, 본사정원수, 본사휴가자수, 본사출장자수, 본사시간외 근무명령서 승인건수, 현본사소속재택근무자수, 조식메뉴, 중식메뉴, 석식메뉴, 중식계, 석식계

b. **train.csv:**
일자, 요일, 본사정원수, 본사휴가자수, 본사출장자수, 본사시간외 근무명령서 승인건수, 현본사소속재택근무자수, 조식메뉴, 중식메뉴, 석식메뉴

c. **sample_submission.csv:**
일자, 중식계, 석식계

(1) 참고 코드**: LGBM**을 활용한 예측 **Baseline +EDA - DACON**

❖ 코드흐름

1. 패키지 임포트
2. 데이터 분석, 전처리

- 결측값 확인, min/max 이상한 칼럼 확인
- object columns / numerical columns 분리
- 일자 칼럼-> 년/월/일 분리 dates.str.get(0)

3. EDA (Exploratory Data Analysis) & VIZ

- 중식계, 석식계 histplot 확인: 타겟값 골고루 분포됨. 두 plot의 x축의 값이 다름
- 석식계 0이 40건 이상: 자기계발의 날, 가정의 날 등
- 본사 정원수 distplot, 본사정원수와 중식/석식계 간 scatterplot 확인
  : 본사정원수 증가, 정원수와 중식계 석식계간 유의한 상관은 없음
- 연 별 본사정원 수 lineplot : 대략 300명 인원 충원
- 본사휴가자 수 displot, 본사휴가자 수와 중식계/석식계 간 scatterplot
  : 본사휴가자 수와 중식계, 석식계 간 음의 관계 확인
- 본사 시간 외 근무 명령서 승인건수 displot, 중식계/석식계 간 scatterplot
  : 양의 관계. 추가 근무 있는 날에 챙겨먹는 경향
- 현본사 소속 재택근무자수 displot, 중식계/석식계 간 scatterplot
  : 큰 관계 없음. 코로나 영향으로 2019 이후 재택근무자 수 증가

4. 피처 엔지니어링

- 휴가자 비율, 출장자 비율, 추가근무자 비율, 휴가자 비율, 출장자 비율, 추가근무자비율 추가
- heatmap 확인

5. Modeling

- object 칼럼에 대한 라벨인코딩
- y1, y2 분리
- LGBMRegressor 이용
- score, MAE(mean_absolute_error) 확인

6. Submission

- submission에 중식계, 석식계 칼럼 추가

(2) 참고 코드**: [Q Branch, Private 3위] XGBOOST와 원초적 본능 - DACON**

❖ 코드흐름

1. df.corr(), sns.pairplot(df)로 상관관계 확인
2. 요일 별 중식계/석식계 확인
    a. 중식은 월요일이 가장 많고 금요일로 갈수록 줄어든다.
    b. 석식은 수요일이 가장 적음.
3. 중식계 / 석식계 sns.kdeplot(x) 확인 : 정규분포를 띄므로 정규화 필요 없음
4. 데이터 정리
    a. 칼럼명 영어로 바꿈
    b. 날짜 datetime타입의 월, 별, 일로 구분, 아침 칼럼은 삭제
    c. 월별 점심/저녁 사람 수 확인 > 연말이 적음
5. 메뉴 구분
    a. 스페이스로 메뉴 별 list로 저장 (쌀밥은 모두 밥으로 통일)
    b. 메인, 국, 반찬 나눠주기 (누락 확인)
    c. 가장 많이 나온 메뉴 확인
    d. 빈 값은 'None'으로 채우기

6. train / test 데이터 생성
    a. lunch_train 생성, 인코딩
    b. corr() 확인: 밥, 국, 메인반찬이 약간의 상관관계 보임
    c. dinner_train - None 값 있는 행 삭제, 인코딩 astype('category'), cat.codes
    d. test 생성

7. 다 합친 train / test 데이터 생성

8. 학습: 점심메뉴로만 구성해서 예측 (점심메뉴가 저녁까지 관계있는지 확인)
    a. XGBoost 사용
    b. max_depth, n_estimators, colsample_bytree, colsample_bylevel 파라미터에 대한 GridSearchCV 진행
    c. 최적 파라미터로 학습, 예측 진행
    d. 제출 파일 생성
9. 학습2: 점심 저녁 메뉴 별(따로) 예측
    a. 마찬가지로 진행
    b. 제출파일 생성

❖ 배운점

- EDA로 전 칼럼 별 타겟변수와의 관계를 시각화해서 파악한다. 한번에 2개 이상의 변수와 타겟을 시각화해보면 좋을 것 같다.

- 비율 변수를 생성해 활용할 수 있다.
- 모든 변수를 이용해 y1, y2 타켓 따로 train 하는 방법과 변수도 구분해 따로 타켓 train하는 방법이 있다.
- kdeplot을 확인해 정규분포 모양을 보이면 정규화를 할 필요가 없다.
- 카테고리화 하여 다양한 문자형을 인코딩 할 수 있다. 이를 통해 상관관계를 파악하는 것이 꽤 유용해 보인다.
- EDA를 자세히 한 후 훈련을 진행한다.