3주차 결과 보고서

전공: 컴퓨터공학 학년: 2학년 학번: 20221559 이름: 박준우

1. XGBClassifior을 이용했다.

모델 선택의 이유로는 비선형적일 수 있는 데이터에 적합하며, 사용이 어렵지 않고 강력한 성능을 제공해 줄 수 있었기 때문이다. 사용하며 성능을 개선할 방법을 찾던 중 찾은 것인데, One-Hot Encode 하지 않고 범주형 데이터를 다룰 수도 있었다.

2. 대분류 중 결측값이 있는 열 삭제, 및 소분류 전체 삭제

대분류는 1부터 4까지의 분포가 거의 고르게 나타나고, 31개밖에 없어서 최빈값 등으로 채우는 것이 적합하지 않다고 판단하여 결측치가 포함된 열을 삭제했다.

소분류는 총 217개의 많은 범주를 가지고 있고, 결측값의 수도 상대적으로 많아서 (가장 많이 나온 분류보다 2배 가량 많다) 모델 성능에 부정적으로 작용 우려하여, 해당 열을 삭제했다. 소분류를 삭제한 후 모델 성능에 도움이 되었다고 생각한다.

3. 실제 클릭하는 상황을 고려한 후 최대한 영향이 있을 것 같은 feature들로 구성해 만들어 봤다.

1) order\_ratio

한 세션 내에서 어느 정도의 구간에 속해있는지를 나타냈다. 같은 session ID를 가진 값들의 개수를 세서, order를 이로 나눈 후, 4를 곱하고 정수로 내림했다. 이로 4개의 구간 정보로 나뉘어 numerical feature로 사용했다. Baseline 코드에 order가 포함되어 있지 않았는데, 이것과, order 자체도 포함하니 성능 개선이 많이 되었다.

2) location에서 1~6까지 나뉘어 있는데, feature의 특성을 잘 반영하지 못 하는 것 같아서 top / bottom, left / middle / right 구분하는 두개의 feature로 구분했다. 이들 각각은 categorial feature로 사용했다. 적용 후 성능 개선에 도움이 되었다.

3) country

29, 9, 24 코드를 제외하고는 전체 비율이 눈에 띄게 적었다. 따라서 29, 9, 24에 대해서만 one-hot encode 하여 진행하였다. 이것과 하이퍼파라미터에 categorial feature이라고 전달한 것과 차이는 작았다.

4. hyper-parameter tuning은 진행해서 모델을 최적화하겠다는 것보다 모델이 새로운 데이터에 대해 제 성능을 낼 수 있도록 해주는 것 같았다. GridSearch는 진행할 때, 각 파라미터를 세세하게 넣으면 너무 시간이 오래 걸려서 Optuna 사용하는 것이 좋았고, 적절한 조합을 찾았다고 생각한다.

5. public에서 제출한 것 중 가장 점수가 높았던 것, feature engineering이 잘 되었던 것 같았던 것으로 2개를 골라서 private 결과로 냈는데, private 결과 점수가 public 점수에 비해 많이 높았다.

public 전체 25위에서 private 공동 1위가 되었는데, private 결과가 좋은 사람들의 제출 횟수가 적은 것으로 보아, public 점수를 높이 가져가려던 시도가 과적합으로 이어졌던 것이 아닌가 생각이 든다.