1. 주제 및 데이터

신용카드 고객 세그먼트 분류를 주제로 한 데이콘 대회

2018년 7월부터 12월까지 고객별 금융활동 데이터

- train data: 회원 정보 + segment, 신용정보, 승인매출 정보, 청구입금 정보, 잔액 정보, 채널 정보, 마케팅 정보, 성과정보

- test data: 회원정보, 신용정보, 승인매출 정보, 청구입금 정보, 잔액 정보, 채널 정보, 마케팅 정보, 성과정보

2. 코드 리뷰

1) EDA

segment 비율을 살펴보면 E가 약 80%를 차지하고, A/B는 매우 적은 불균형 데이터이다.

boxplot, heatmap을 통해 변수간의 상관관계를 파악하였다.

2) 전처리

(1) 결측치 처리: 여러 변수에 결측치가 존재하므로 결측치가 너무 많은 변수는 삭제하였고, 변수 특성에 따라 '기타'로 분류하거나 '중앙값'으로 대체하였다.

(2) 자료형 변환: 문자열 변수를 숫자로 인코딩하였다.

(3) 파생변수 생성: 청구서수령방법이 당사멤버십인 고객의 등급은 C,D,E만 존재하므로 당사멤버십 여부 (0,1)로 파생변수를 생성하였다.

(4) 변수 삭제: 파생변수에 대한 원천변수를 삭제하고, 모든 값이 동일한 변수를 삭제하였다.

3) 모델링

XGBoost, LightGBM, CatBoost 모델을 각각 학습시켜 성능을 개선하고, 세 모델을 앙상블하여 최종 모델을 완성했다.

(1) 전체 변수 수가 너무 많아 학습 속도 저하 및 과적합 우려가 있었으나 XGBoost, CatBoost, LightGBM으로 각각의 feature importance를 파악해 상위 변수를 추출했다.

(2) A/B 클래스 데이터가 부족했기 때문에 SMOTE+클래스 가중치를 적용하여 오버샘플링을 수행하였다.

(3) XGBoost, CatBoost, LightGBM 하이퍼파라미터 튜닝을 진행하였다.

(4) soft voting 방식으로 모델별 확률을 가중 평균하여 안정적이고 높은 성능을 확보하였다.

3. 차별점 및 배울 점

EDA를 통해 데이터의 불균형을 확인하고 수가 적은 데이터 유형에만 오버 샘플링을 적용한다는 점이 인상 깊었다. 또 단순히 XGBoost, CatBoost, LightGBM 모델을 앙상블하는 것이 아니라 soft voting을 통해 모델별 가중치 비율을 조절하면서 더 안정적이고 높은 성능의 모델을 도출해 낼 수 있다는 것을 배웠다.