2024-11-18 수상작 리뷰

고객 대출등급 분류 AI 해커톤

https://dacon.io/competitions/official/236214/codeshare/9679?page=1&dtype=recent

데이터

- **train.csv [파일]**
  ID : 대출 고객의 고유 ID
  고객 관련 금융 정보
  대출금액
  대출기간
  근로기간
  주택보유상태
  연간소득
  부채 대비 소득 비율
  총계좌수
  대출목적
  최근 2년간 연체 횟수
  총상환원금
  총상환이자
  총연체금액
  연체계좌수
  대출등급 : 예측 목표

- **test.csv [파일]**
  고객 관련 금융 정보
  ID : 대출 고객의 고유 ID
  대출등급이 존재하지 않음

- **sample_submission.csv [파일] - 제출 양식**
  ID : 대출 고객의 고유 ID
  대출등급 : test.csv에서 제공된 고객의 대출등급을 예측하여 기입

코드 흐름

1) **EDA**
   ID drop
2) **Feature Engineering & Processing**
   대출기간 칼럼 전처리

LabelEncoder로 범주형 변수 인코딩
파생변수 생성: EDA를 통해 '총상환원금', '총상환이자'가 주요한 영향을 미치는 것을 파악했고, 두 변수의 결합('총상환원금'/'총상환이자')으로 파생변수를 생성한다. 또한, '총상환원금'/'대출금액'으로 '상환비율' 변수이라는 변수도 생성한다.

변수 제거 (RFECV, Feature Importance)
Feature Importance를 찍어보니 값이 0에 가까운 변수들이 몇 개 있기 때문에 그것을 아예 제거한다. 변수 선택법인 RFECV를 활용해서 변수를 3개로 줄인다. ('대출기간', '총상환원금/총상환이자', '상환비율')

3) 모델 학습(Linear Regression)

하이퍼파라미터 튜닝 (Optuna): ET, RF, DT, XGB 총 4가지 모델들에 대해 튜닝을 진행한다. XGB의 경우 tree_method에 따라 점수가 많이 달라지는 것을 확인해 튜닝 작업에 추가한다.

앙상블 모델 (Stacking): Voting과 Stacking, 그리고 여러 모델끼리 조합을 해보면서 가장 성능이 좋은 조합을 찾아본다. Stacking(ET+XGB+DT+RF)이 제일 좋다.

4) 후처리

예측값 할당

차별점, 배울점

이 수상작에서는 RFECV를 이용하여 변수를 제거하였다. RFECV는 재귀적 특성 제거와 교차 검증을 결합한 변수 선택 방법으로, 중요도가 낮은 변수를 반복적으로 제거하는 과정을 통해 최적의 변수 조합을 찾아낸다. RFECV는 먼저 모든 변수를 사용하여 모델을 학습시킨 다음에, feature importance나 계수를 기반으로 변수의 중요도를 평가해서 가장 중요도가 낮은 변수를 제거하고 다시 모델을 학습시킨다. 이렇게 만들어진 변수 조합들에 대해 교차 검증을 수행하여 모델을 평가한다. 이런 과정을 계속 반복해서 성능이 가장 높은 변수 조합을 선택하는 과정을 수행한다.

이 수상작에서는 파생 변수를 만들고, RFECV를 수행하여 모델의 성능을 높일 수 있는 변수를 찾는 것에 치중하였다.