랜덤포레스트 모델을 이용한 카드사 고객 세그먼트 분석

금융 | scikit-learn | imbalanced-learn | scipy.stats | Python

프로젝트 개요

데이터 개요

- 카드사 고객별 카드 사용 데이터(총 거래금액, 리볼빙 잔액 등), 인구통계학적 데이터(나이, 교육 수준, 결혼 상태 등), 금융 데이터(신용한도 등) 그리고 이탈 여부를 Yes/No로 나타내는 field로 구성됨.

분석 목표

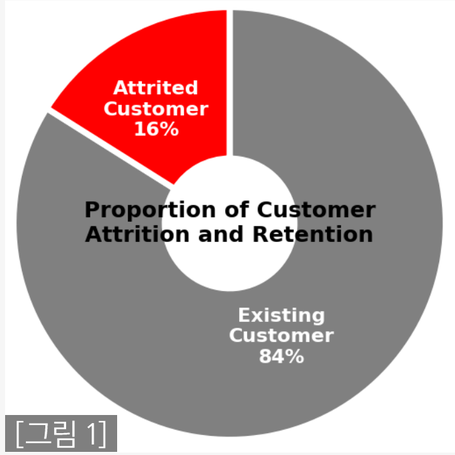
- 이탈 여부가 포함된 데이터인만큼 머신러닝 모델을 활용해 고객의 이탈 여부를 예측할 수 있는 모델을 도출하고자 함.
- 기본적으로 모든 변수를 포함시켜 모델을 학습시키는 접근부터 시작했으나 너무나 상식적인 결과 혹은 인과가 뒤집힌 결과들이 포함되어 있어 유의미한 메세지를 도출하기 어려웠음.
  (e.g., '총 거래금액이나 총 거래횟수가 낮을수록 이탈 확률이 높음' → 상식적인 결과, '최근 6개월 간 카드사로부터 연락 받은 횟수가 높을수록 이탈 확률이 높음' → 이탈 조짐이 보여 카드사가 연락을 취한 것일 가능성 높음, 즉 인과가 뒤집힌 결과로 해석 가능)
- 또한 모든 변수를 활용한 모델은 자사 카드 사용에 대한 데이터를 많이 포함해 신규 고객 유입 시는 적용이 어렵다는 한계도 존재함.
- 따라서 보다 **(1) 상식적 유추가 어려운 메시지를 함축하며 (2) 자사 카드 사용 기록이 없는 신규 고객에도 적용 가능한 모델을 도출**하는 것을 분석 목표로 잡음. 또한 도출한 모델을 원본 데이터셋에 적용함으로써 **이탈 조짐을 보이는 고객을 예측**하는 부가적인 효과도 노림.

랜덤포레스트 모델을 이용한 카드사 고객 세그먼트 분석

랜덤포레스트 알고리즘 기반 모델 학습

변수 선별

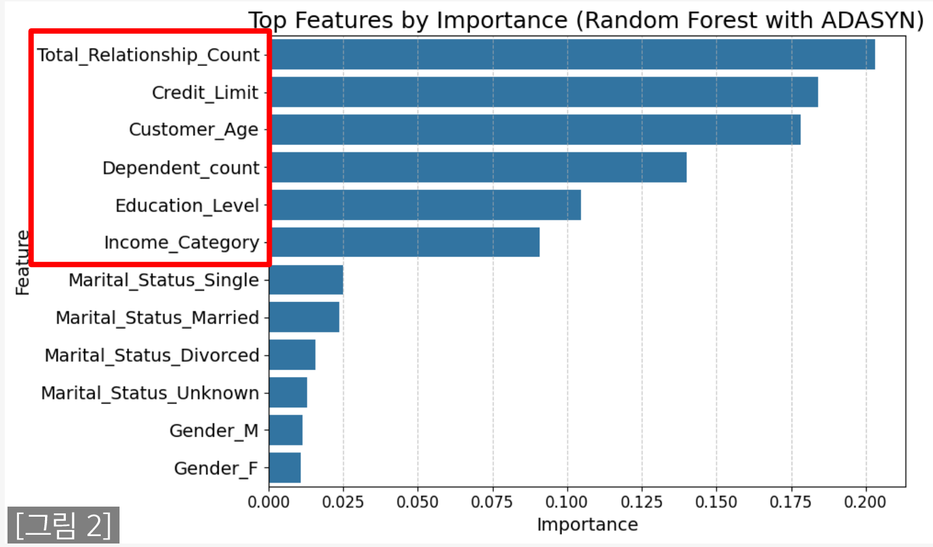
- 자사 카드 사용 데이터가 없는 신규 고객 유입 시에도 수집 가능한 데이터는 아래와 같이 총 8가지가 존재함.
  - (수치형 변수) 나이, 부양가족 수, 자사 금융상품 이용 개수, 신용한도
  - (범주형 변수) 성별, 혼인 상태, 교육 수준, 소득 수준
    (타 카드사 이용 행태 정보는 알 수 없다고 가정)
- 범주형 변수 중 위계가 존재하는 교육 수준과 소득 수준에는 Ordinal Encoding을, 그렇지 않은 성별과 혼인 상태에는 One-Hot Encoding을 적용함.



[그림 1]

랜덤포레스트 모델 학습

- 이탈 고객과 잔존 고객 데이터 수에 비대칭이 존재하는 바([그림 1]) 전처리 과정에서 오버샘플링 기법인 ADASYN을 적용함. (sampling strategy = minority)
- 모델에 의해 도출된 Feature Importance는 (1) 자사 금융상품 이용 개수 (2) 신용 한도 (3) 나이 (4) 부양가족 수 (5) 교육 수준 (6) 소득 수준 순으로 높았으며 혼인 상태와 성별은 상대적으로 중요도가 떨어지는 양상을 보였음. ([그림 2])



[그림 2]

랜덤포레스트 모델을 이용한 카드사 고객 세그먼트 분석

랜덤포레스트 예측 결과 기반 고객 그룹 분석

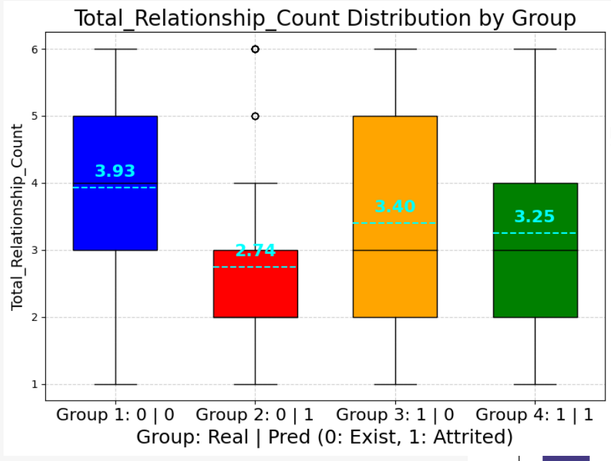
그룹 설정

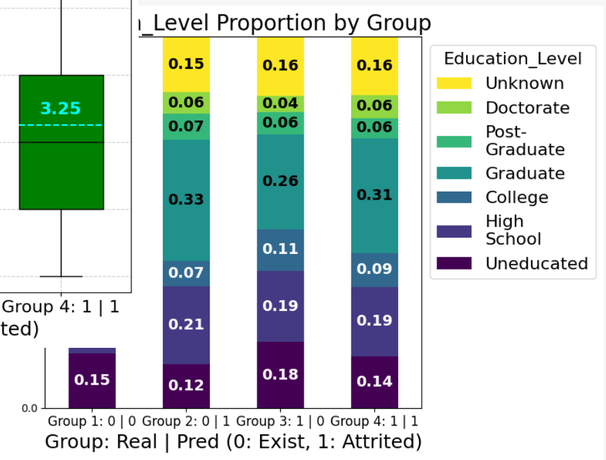
- 도출 모델을 원본 데이터셋에 적용한 뒤 예측 결과와 실제 값을 기반으로 표와 같이 그룹을 분류함.
- 그룹 1 (모델 그룹): 안정적인 이용행태를 보이는 그룹으로서 신규 고객 진입 시 이와 유사한 특성을 지니는 경우 안전 고객이라 판단하는 롤모델로 삼을 수 있을 것으로 판단.
- 그룹 2 (위험 그룹): 잔류 고객이지만 모델이 이탈로 예측한 그룹으로 이탈 예방 조치가 필요한 대상으로 판단.

그룹	실제	예측
그룹1	잔류	잔류
그룹2	잔류	이탈
그룹3	이탈	잔류
그룹4	이탈	이탈

그룹 특성 분석 방법

- 도출한 모델을 적용하면 이탈 예측의 결과를 알 수 있긴 하지만 모델을 적용하지 않고도 보다 직관적으로 활용할 수 있는 지표를 도출하고자 함. (e.g., '계열사 상품을 많이 이용하는 고객은 추후 안정적인 이용 행태를 보일 가능성이 높음')
- 이를 위해 그룹별로 모델에 활용한 변수들에 대한 시각화를 진행함.
  - 수치형 변수: Boxplot with Mean Line
  - 범주형 변수: Stacked Bar Chart





랜덤포레스트 모델을 이용한 카드사 고객 세그먼트 분석

랜덤포레스트 예측 결과 기반 고객 그룹 분석

수치형 변수 분석 결과

그룹 1 (모델 그룹)

- 4그룹 중 **가장 높은 평균 자사 금융상품 이용 수**와 **가장 낮은 평균 부양가족 수**를 보임.([그림 3], [그림 6])
- 그룹3 제외 시 가장 **높은 평균 신용한도**와 가장 **낮은 평균 나이**를 보임.([그림 4], [그림 5]) 이는 '잔존 그룹(그룹1, 그룹2) 중 비교적 높은 신용한도와 비교적 낮은 평균 나이를 보이는 군집'으로도 해석 가능함.

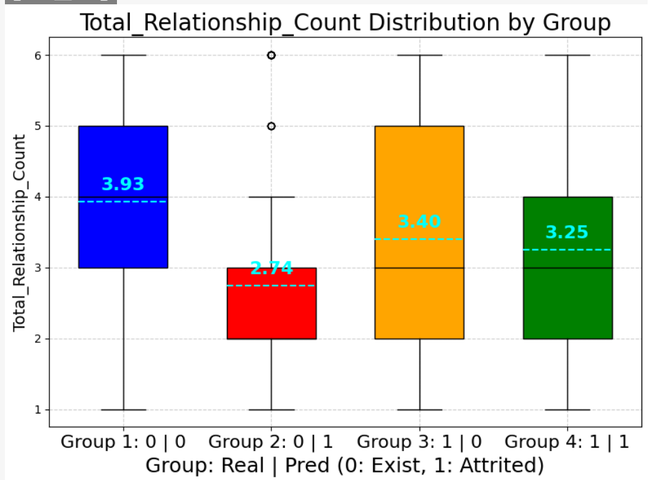
→ 신규 고객 평가 시 이와 유사한 특성을 지닌 고객은 안정적 이용행태를 보일 것이라고 직관적인 판단을 내릴 수 있음.

그룹 2 (위험 그룹)

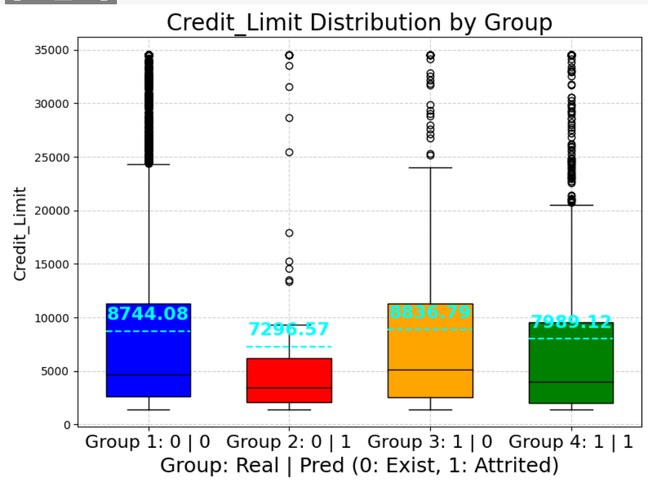
- 모든 지표에서 4그룹 중 가장 극단적인 수치를 보임.(가장 높은 평균 나이와 평균 부양가족 수 ([그림 5], [그림 6]), 가장 낮은 평균 자사 금융상품 이용 수와 신용 한도 ([그림 3], [그림 4]))

→ 잔존 고객 중 이러한 특성을 보이는 고객들에 대한 이탈 예방책 적용이 필요함.

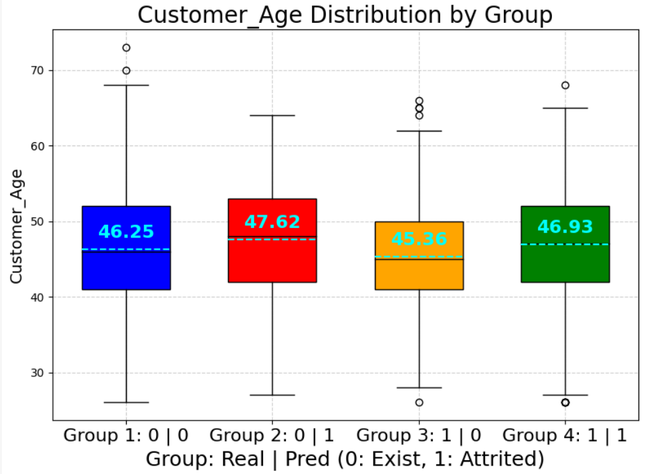
[그림 3]



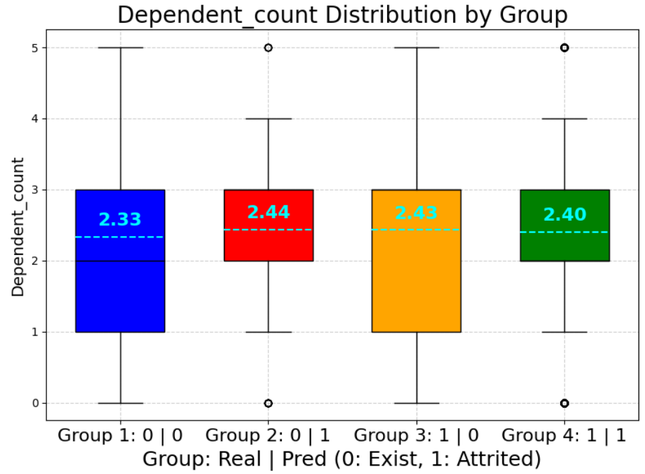
[그림 4]



[그림 5]



[그림 6]



랜덤포레스트 모델을 이용한 카드사 고객 세그먼트 분석

랜덤포레스트 예측 결과 기반 고객 그룹 분석

범주형 변수 분석 결과

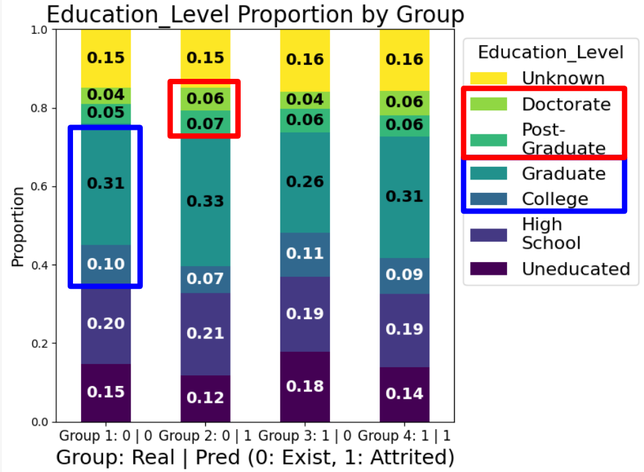
그룹 1 (모델 그룹)

- 4그룹 중 **중위 교육수준(대졸+대학생)**과 **중위 소득 및 중위~고소득군(60~80K, 60~120K)**에 속하는 비중이 가장 높음. ([그림 7], [그림 8])
- 교육수준이나 소득수준의 양극단(초상위/하위 학력, 초고/저소득)에서는 평균적인 비율을 보임.

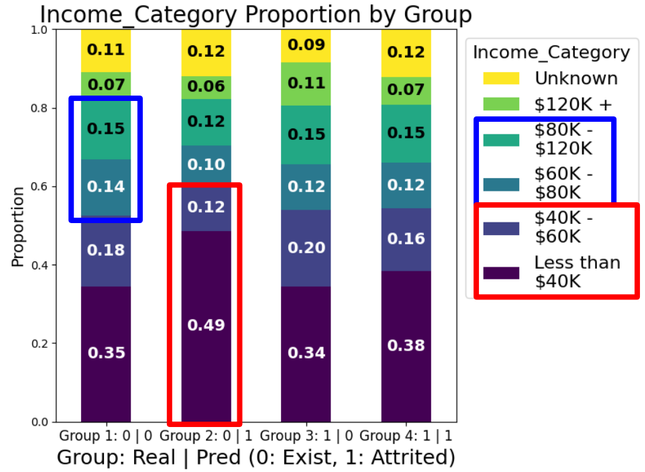
그룹 2 (위험 그룹)

- 교육수준에서는 4그룹 중 **초상위 학력(박사졸+석사졸)**에 속하는 비중은 가장 높고, 상대적 하위 학력(무학력+고졸)에 속하는 비중은 가장 낮음. 즉 보편적으로 높은 교육 수준을 보유한 것으로 해석 가능. ([그림 7])
- 소득수준에서는 4그룹 중 초고소득군 및 상대적 고소득군 (>120K, >80K)에 속하는 비중은 가장 낮고, **초저소득군 및 상대적 저소득군(<40K, <60K)**에 속하는 비중은 가장 높음. 즉 보편적으로 낮은 소득 수준을 보유한 것으로 해석 가능. ([그림 8])

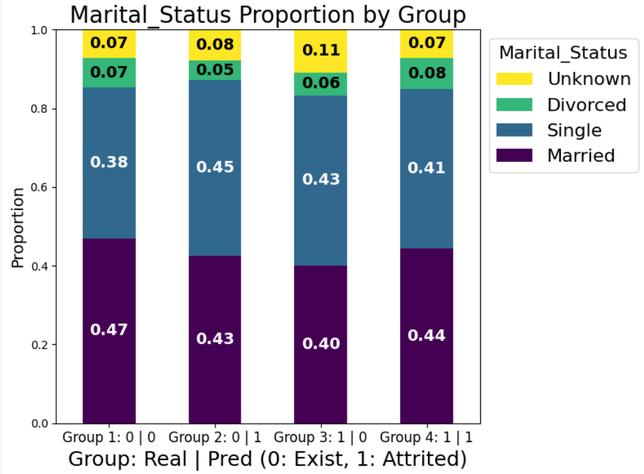
[그림 7]



[그림 8]



[그림 9]



[그림 10]