

인공지능 데이터 전처리 결과서

은행 고객 이탈 예측 프로젝트

1. 프로젝트 개요

- 프로젝트명: 은행 고객 이탈 예측 모델 개발
- 목표: 고객 특성과 거래 정보를 바탕으로 이탈 여부 예측
- 데이터 출처: 깃허브
 - https://github.com/adin786/bank_churn/tree/main

2. 데이터 설명

- BankChurners.xlsx
- 데이터 수: 10127
- 변수 개수: 21

3. 전처리 요약

1. 결측치 처리
 - A. 데이터 결측치 없음
2. 피처 엔지니어링
 - A. 원 - 핫 인코딩
 - ① 결혼여부 - 기혼, 미혼, 이혼, Unknown
 - ② 카드등급 - Blue, Gold, Silver, Platinum
 - ③ 성별 - Male, Female

B. 데이터 구간화

① 나이

나이	나이 구간화
0 ~ 30	0
30 ~ 40	1
40 ~ 50	2
50 ~ 60	3
60 ~ 100	4

② 카드보유기간

카드 보유 기간(월 단위)	카드 보유 기간(년 단위)
0 ~ 24	1
24 ~ 36	2
36 ~ 48	3
48 ~	4

③ 수입 구간

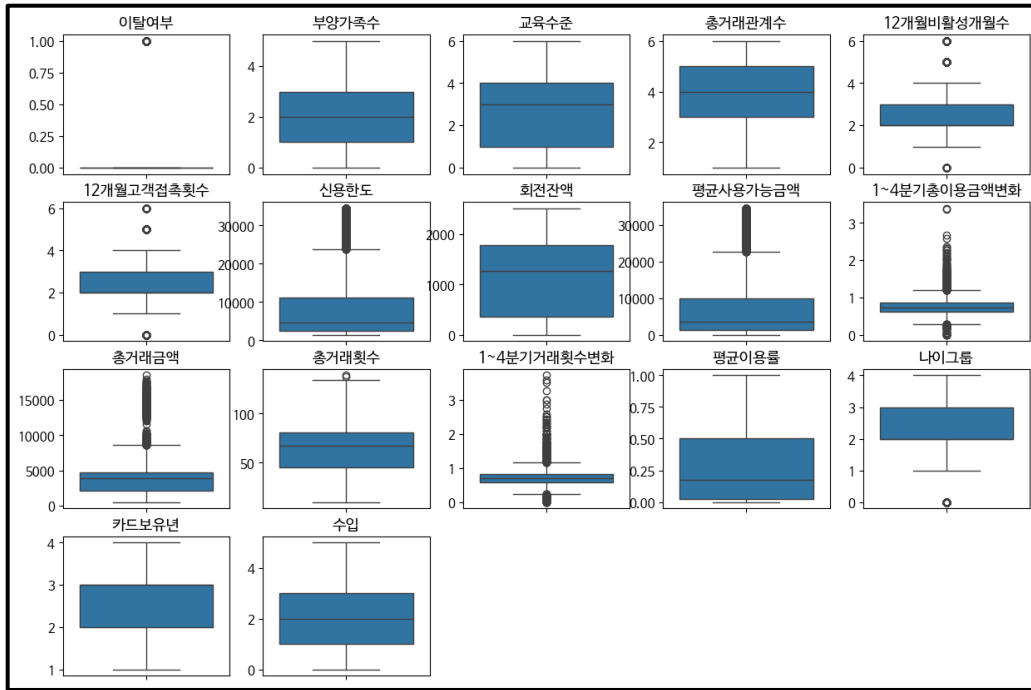
수입 구간	수입 구간 범주화
Unknown	0
Less than \$40K	1
\$40K - \$60K	2
\$60K - \$80K	3
\$80K - \$120K	4
\$120K ~	5

④ 교육 수준

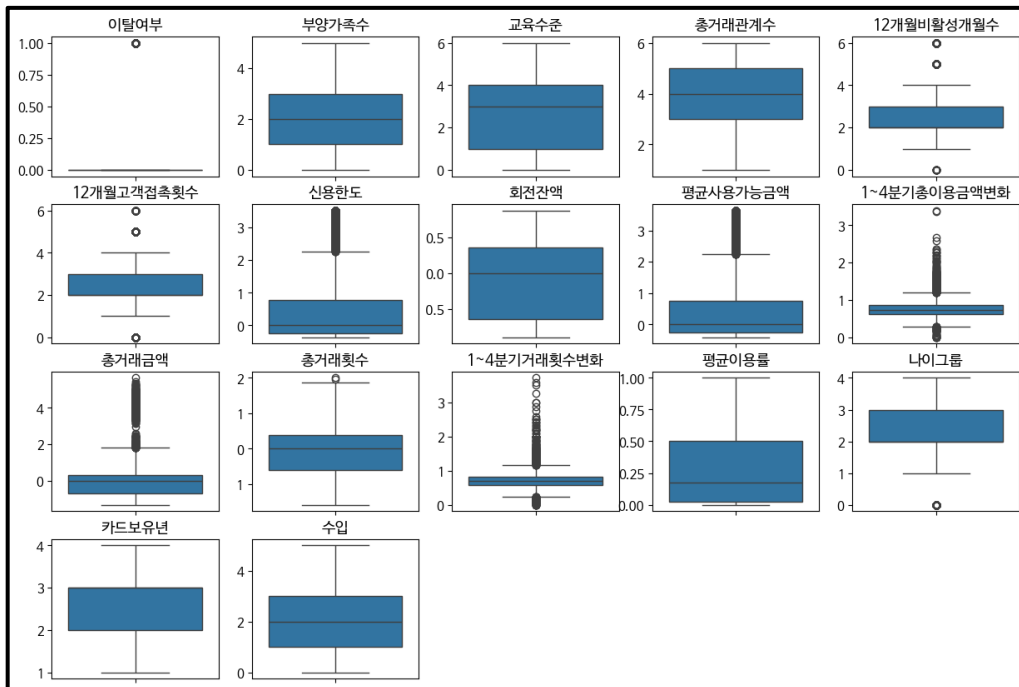
교육 수준	교육 수준 범주화
Unknown	0
Uneducated	1
High School	2
College	3
Graduate	4
Post-Graduate	5
Doctorate	6

3. 이상치 탐색 및 스케일링

A. 이상치 탐색

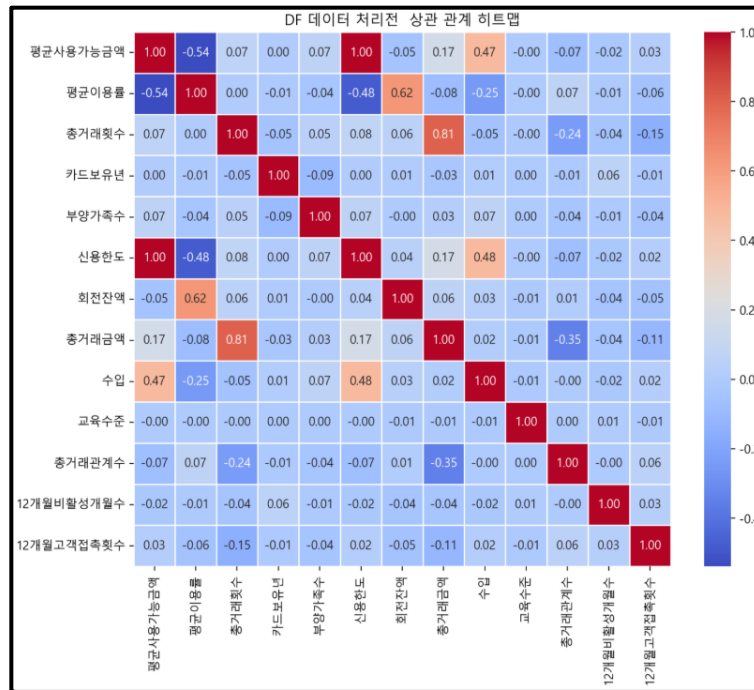


B. Robust 스케일링

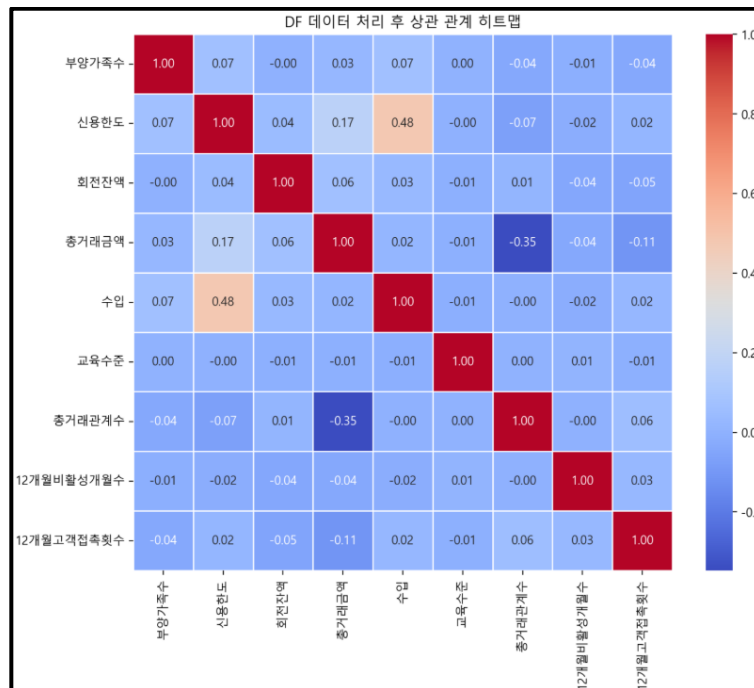


4. 상관관계

A. 처리 전 상관관계

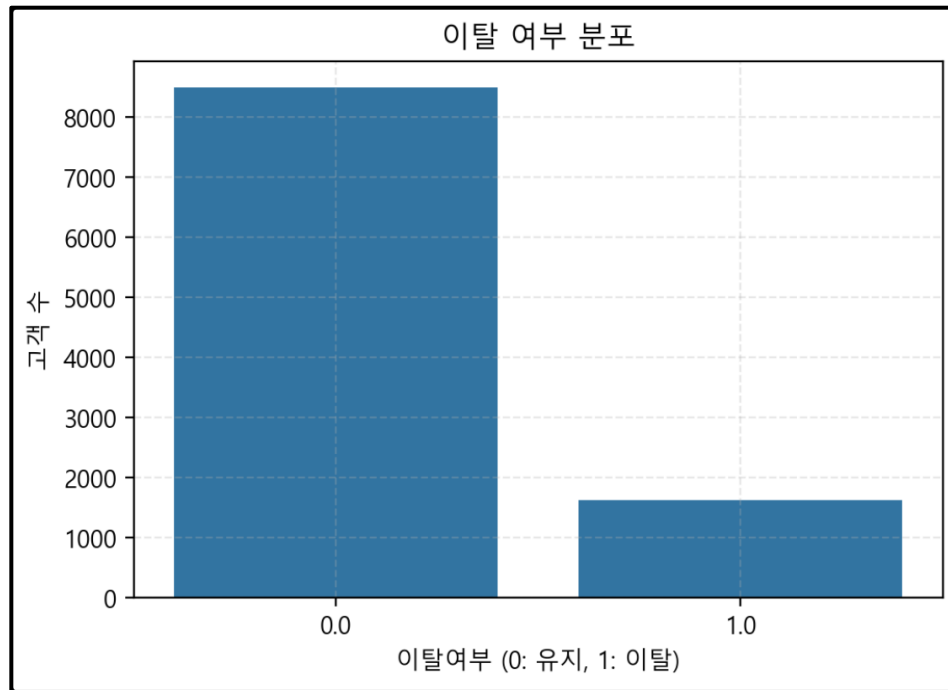


B. 상관성 높은 변수 제거 후 상관관계



5. 클래스 불균형

A. 원본 데이터



B. SMOTE 을 진행한 오버 샘플링

