인공지능 데이터 전처리 결과서

팀: 1조(김나현, 문창교, 이경현, 이승규, 정래원)

프로젝트명: 고객 이탈 예측 (Customer Churn Prediction)

데이터 출처: Kaggle – Bank Customer Churn Dataset

데이터 규모: 총 10,000개 행, 12개 열

1. 데이터셋 로드 및 기본 전처리

1-1) 데이터 로드 및 기본 정보

본 분석은 Kaggle에서 제공하는 Bank Customer Churn Dataset을 활용하였다.

총 10,000개의 고객 데이터와 12개의 변수로 구성되어 있으며, 고객의 인구통계학적 특성, 계좌 관련 정보, 신용 점수 및 은행 상품 이용 현황 등이 포함되어 있다.

1-2) 결측치 탐색 및 초기 전처리

EDA 전, 기본 전처리로 다음을 수행했다.

- 고유 ID 컬럼(`customer_id`) 삭제: 모델링에 불필요하고 예측과 무관한 식별자.
- 범주형 변수 인코딩: `gender`, `country` 변수에 대해 `LabelEncoder` 적용.

이를 통해 분석에 불필요한 변수는 제거, 범주형 변수는 수치형으로 변환해 EDA 전, 데이터셋에 대해 기본적인 전처리를 완료했다.

2. 결측치 및 이상치 처리

2-1) 결측치 처리

`df.isnull().sum()`을 통해 결측치를 확인한 결과, 결측값은 존재하지 않았다.

2-2) 이상치 탐색 기준

연속형 변수는 사분위수를 활용한 **IQR** 기반 탐색을 실시했다.

(하한(Q1 - 1.5×IQR) 미만 or 상한(Q3 + 1.5×IQR) 초과 값 → 이상치로 판단)

범주형 변수는 각 범주의 비율을 확인해 1% 미만의 희귀 카테고리를 이상치로 간주했다.

2-3) 이상치 탐색 결과

변수	이상치 수	비율	처리 방법
age	359	3.59%	행 삭제
credit_score	15	0.15%	행 삭제
products_number (카테고리 4)	60	0.006%	카테고리 3과 통합

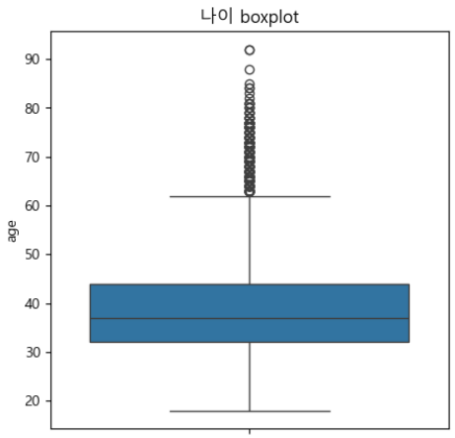
age와 credit_score의 이상치는 겹치지 않아 총 374개 행(3.74%) 을 삭제했다. 또한 products_number 변수에서 4개 상품 보유 고객(60명)은 극소수로 확인되어, 인접한 그룹(3개 상품 보유 고객, 266명)과 통합하여 카테고리 균형을 확보했다.

3. EDA

EDA(Exploratory Data Analysis, 탐색적 데이터 분석)는 단변수 분석 ⇒ 이변수 분석 ⇒ 다변수 분석 순으로 진행했다.

3-1) 단변수 분석

단변수 분석의 목적은 각 변수의 개별 분포를 파악하는 것이다. 변수마다 .describe()으로 기술통계량을 확인하고, 시각화(히스토그램, 박스플롯, countplot)를 진행했다. 분량상 보고서에는 진행한 시각화 가운데 대표 그래프 1개를 싣도록 하겠다.



boxplot 사용/ age 변수의 분포 확인(단변수 분석)

이를 통해 각 변수의 스케일/이상값의 존재 여부/왜도 등을 확인할 수 있었다. 단변수 분석 과정에서 도출한 주요 인사이트는 EDA.ipynb 파일 및 아래 표에서 확인할 수 있다.

표 1.1 단변수 분석: 연속형 변수

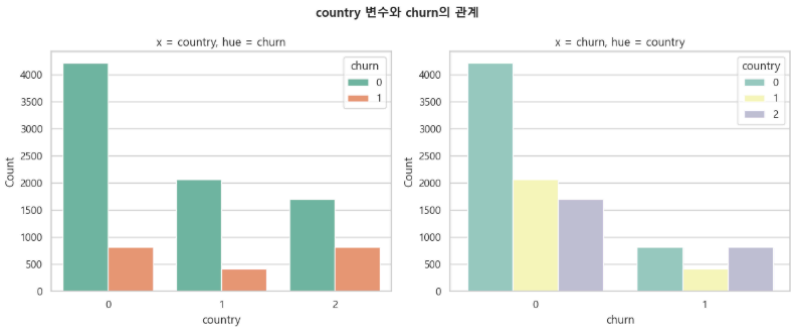
변수명	단변수 분석 인사이트
credit_score	- 신용 점수는 대체로 정규분포를 따르나, 400이하의 값에서 일부 이상치가 보임 - 신용 점수가 높아질수록 사람 수가 적어지다가, 갑자기 증가하는 구간이 보임(800~850)
age	- 50%의 데이터가 (32~44)세에 밀집돼 있음 - 고령층의 수는 적음 - 이상치가 많음(359개) ⇒ boxplot에서 62세 이상은 전부 이상치로 판단됨
balance	-histplot을 보니, 잔고가 0인 고객이 매우 많음 -잔고가 0인 고객이 전체 고객의 36.5%
estimated_salary	-Q2가 상자의 정중앙에 오고, 위 아래 수염 길이도 비슷함. -연봉 변수는 데이터가 대칭/고르게 분포돼있음.

표 1.2 단변수 분석: 범주형 변수

컬럼명	단변수 분석 인사이트
products_number(1~4)	- 1개(50.84%)나 2개(45.9%)의 상품을 이용하는 고객이 많음 - 4개 상품 이용 고객은 매우 적음
country(0~2)	- 프랑스 50.1%, 스페인 24.8%, 독일 25.1% - 고객층의 50%가 프랑스
gender(0/1)	-남자 54.6%, 여자 45.4%로 성비 비슷
credit_card(0/1)	-보유 70.5%, 미보유 29.5%
active_member(0/1)	- 활동적인 회원 51.5%, 비활동적인 회원 48.5%로 비슷함
tenure(0~10)	-0년(처음 가입한 고객)과 10년은 400명대 -나머지 기간(1년~6년)은 대체로 비슷한 수준임(900~1000명대)

3-2) 이변수 분석

이변수 분석은 각 변수와 타겟 변수(churn) 간의 관계 파악이다. 변수 유형에 따라 시각화에 사용하는 그래프 종류를 다르게 했는데, 연속형 변수–박스플롯, 범주형 변수–countplot을 사용했다. 분량상 보고서에는 진행한 시각화 가운데 대표 그래프 1개를 싣도록 하겠다.



countplot사용 / 범주형 변수 country와 타겟 변수 Churn의 관계 (이변수 분석)

해당 과정에서 도출한 주요 인사이트는 EDA.ipynb 파일 및 아래 표에서 확인할 수 있다.

표 2.1 이변수 분석 : 연속형 vs Churn

	타겟 변수 Churn
credit_score	-집단간 신용점수 분포 차이는 없어보임 (이탈집단(`target=1`)과 유지집단(`target=0`)의 중앙값, 사분위수 등 큰 차이는 없음) -다만, 이탈 집단에서 신용점수 400이하의 이상치가 다수 발견됨 (**특이점:** 이탈집단에 **credit_score < 400**인 극단적 저신용 고객들이 집중적으로 존재)
age	- 이탈 집단이 유지 집단보다 연령대가 높음(40중반~50중반). 유지 집단은 상대적으로 젊음(30초~40초)
balance	- 이탈 집단의 잔고 평균이 조금 더 높음 - 유지 집단이 이탈 집단보다 IQR 분포가 *아래쪽으로* 더 넓음 - 유지 고객중 잔고없거나 적은 사람이 많음 - 유지 집단은 잔액이 0인 고객이 많음 => 잔액이 없으면 이탈 가능성 낮음
estimated_salary	-집단간 연봉 분포는 비슷함

표2.2 이변수 분석 : 범주형 vs Churn

	X = Churn	hue = Churn
products_number	- 이탈집단의 이용상품수 1개 >>> 2개 > 3개> 4개 - 유지집단 이용상품수 2개> 1개	- 상품을 3~4개 이용하는 고객 중에선 이탈이 유지보다 월등히 높음 - 상품 1개 이용하는 고객의 이탈율은 약 40% - 상품 2개 이용하는 고객들이 가장 이탈율이 적음
country	- 유지 집단 중에서는 0번 국가(프랑스)의 비율이 가장 높음 - 이탈 집단은 2번(독일)>0번(프랑스)>1번(스페인) 순으로 많음	- 2번 국가(독일)의 이탈율이 절반 정도로 높음
gender	- 유지 집단에선 남자 비율이 높음 - 이탈 집단에선 여자 비율이 높음	- 여자가 남자보다 이탈 비율이 높음
credit_card	- 유지집단/이탈집단 간 신용카드 보유 여부도 비슷함	- 신용카드 보유에 따른 이탈 차이는 없음
active_member	- 이탈 집단일수록 비활동회원이 많음 - 유지 집단일수록 활동회원이 많음	- 비활동집단일수록 이탈율이 높음
tenure	- 카테고리가 10개로 많아서 그래프상 눈에 띄는 점은 없다.	

3-3) 다변수 분석

다변수 분석이란 3개 이상의 변수 조합에 대해 관계를 분석하는 것이다. 해당 과정에서는 상관관계 분석(heatmap)과 다중공선성(VIF) 검토을 진행했다.

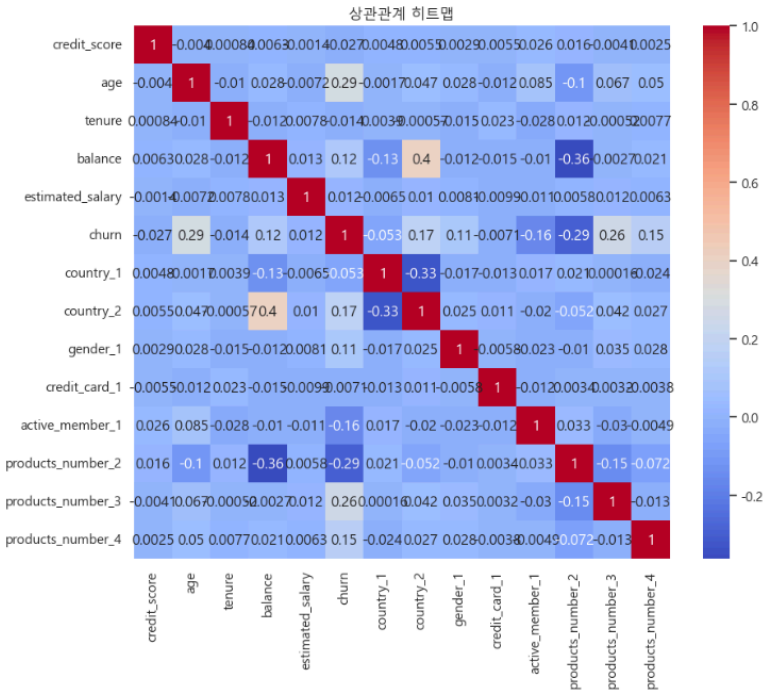
분석 전, Scaling과 Encoding을 선행했다. 연속형 변수의 경우, 모든 변수의 평균이 0, 분산이 1이 되도록 표준화를 진행했다. 범주형 변수는 범주별로 분리해 관찰해야 하므로 One-Hot Encoding을 했다.

(단, 이 스케일링은 EDA 목적의 변환임. 모델링 단계에서는 train/test 분리 후 다시 수행해 데이터 누수를 방지했음.)

3-3-1) 상관관계 분석

피어슨 상관계수(연속형 변수들 사이 상관 정도 측정)과 스피어만 상관계수(순서형 변수들 사이 상관 정도 측정), 두 방법으로 진행했을때, 방법에 따른 유의미한 차이는 없었다(각 방법으로 측정한 결과는 EDA.ipynb 파일에서 확인할 수 있음). 따라서 해당 보고서는 피어슨 상관계수 기준으로 분석한 결과를 담고 있다.

전체 변수를 대상으로 상관관계 히트맵을 그린 결과는 다음과 같다.

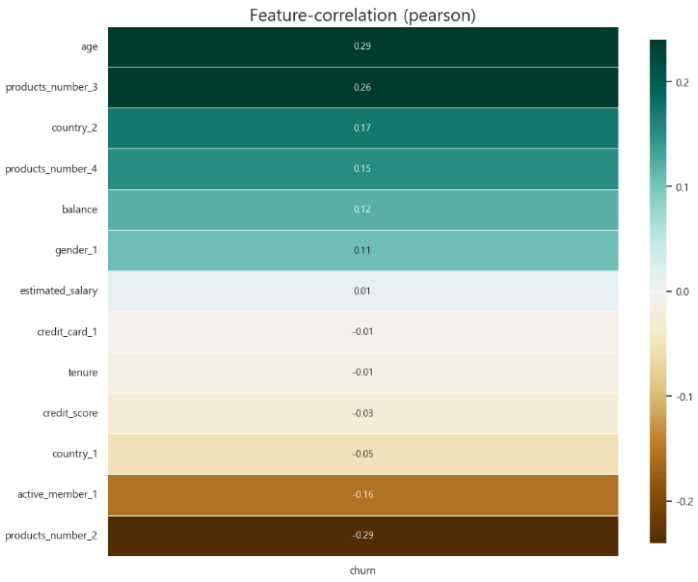


히트맵에서 |상관계수| ≥ 0.3 이면 약한 상관관계가 있다고 판단해 변수 조합을 기록했다. (타겟인 Churn은 제외)

- balance - 상품개수2 (-0.36) : 상품이 2개인 고객일수록 잔액이 낮은 경향(음의 상관)
- balance - country2 (0.4) : country2 지역 고객일수록 잔액이 높은 경향(양의 상관)
- country1 - country2 (-0.33) : 두 지역 더미변수는 음의 상관 => 원핫인코딩 특성상 자연스러운 결과

만약 모델 성능이 목표치까지 올라가지 않을 경우, 약한 상관관계가 있는
balance - 상품개수2 (-0.36) 나 balabce - country2 (0.4) 로 파생 변수를 생성해 추가하는 시도를 할 수 있다.

타겟 변수 `Churn` 과 타 변수의 상관관계 히트맵은 다음과 같다.



(|상관계수| 기준)

상품개수2> 나이 >상품개수3> 국가2 > 활동회원 > 상품개수4 순으로 고객 이탈(타겟 변수 Churn)과의 관련성이 강한 것을 확인할 수 있다.

3-3-2) 다중공선성 검토

다중공선성이란, 특정 변수가 다른 변수들과 강한 선형관계를 가지는 현상으로, 다중공선성이 존재할 경우 모델 해석력 저하를 초래할 수 있다. 이를 정량적으로 평가하기 위해 VIF(Variance Inflation Factor, 분산팽창계수) 지표를 측정했다. 일반적으로 VIF 값이 5 이상이면 다중공선성이 존재한다고 판단하며, 이 경우 변수 제거 또는 PCA(주성분 분석) 등을 통한 차원 축소가 필요하다.

- VIF에 따른 다중공선성 판단 기준

VIF = 1: 다른 변수들과 전혀 상관관계가 없음
1 < VIF < 5: 약한~중간 정도의 상관관계
5 < VIF < 10: 높은 상관관계, 주의 필요
VIF > 10: 심각한 다중공선성, 변수 제거 고려

아래 사진은 VIF 측정 결과이다.

	feature	VIF
0	credit_score	1.001658
1	age	1.110699
2	tenure	1.002220
3	balance	1.401081
4	estimated_salary	1.001055
5	churn	1.352674
6	country_1	1.125197
7	country_2	1.371610
8	gender_1	1.013243
9	credit_card_1	1.001673
10	active_member_1	1.047173
11	products_number_2	1.290294
12	products_number_3	1.087827
13	products_number_4	1.029537

본 데이터셋의 모든 변수에 대해 VIF를 산출한 결과, 모든 변수의 VIF 값이 1점대로 확인되었다.

이는 각 독립변수 간의 상관성이 낮고, 상호 독립성이 확보되어 있음을 의미한다.

따라서 본 프로젝트에서는 변수 제거나 차원 축소(PCA) 등의 추가 조치는 수행하지 않았다.

4. 전처리 과정

4-1) 모델링용 전처리 파이프라인

모델링 단계에서는 `ColumnTransformer` 기반으로 연속형·범주형 변수를 분리하여 처리하였다.

구분	처리 내용
연속형	`StandardScaler` / `MinMaxScaler` (모델 유형별 선택)
범주형	`OneHotEncoder(drop='first', handle_unknown='ignore')`
데이터 분할	`train_test_split(test_size=0.2, stratify=y, random_state=42)`
클래스 불균형	`Churn=1` 비율 약 20.3% → Stratified Split 적용으로 유지
누수 방지	Pipeline 내부에서 `fit`은 train 데이터에만 수행 후 test에는 `transform` 적용