Netflix 고객 이탈률 예측 모델 산출물

I. 데이터 전처리 결과서

I. 데이터 탐색 및 전처리 개요

1. 데이터셋 개요

- **데이터 출처:** Kaggle - Netflix Customer Churn dataset
- **데이터 크기:** 5,000명 고객 × 14개 컬럼
- **데이터 구성**
    - 수치형 변수:
      customer_id, age, watch_hours, last_login_days, monthly_fee, churned, number_of_profiles, avg_watch_time_per_day
    - 범주형 변수:
      gender, subscription_type, region, device, payment_method, favorite_genre
- **데이터 전처리 목적**
    - 데이터 품질 확보
      결측치, 이상치 확인 및 처리로 모델 학습에 적합한 안정적 데이터셋 확보
    - 모델 입력 변수 정제
      불필요 컬럼 제거, 범주형 인코딩, 수치형 스케일링 등을 통해 모델 학습 효율성 향상
    - 특성 간 비교 가능성 확보
      변수 간 단위 차이 제거 및 정규화로 모델 수렴 안정성 강화
    - 예측모델 준비
      전처리 완료된 데이터를 기반으로 고객 이탈 여부(churn) 예측 모델 학습 및 성능 향상 지원

---

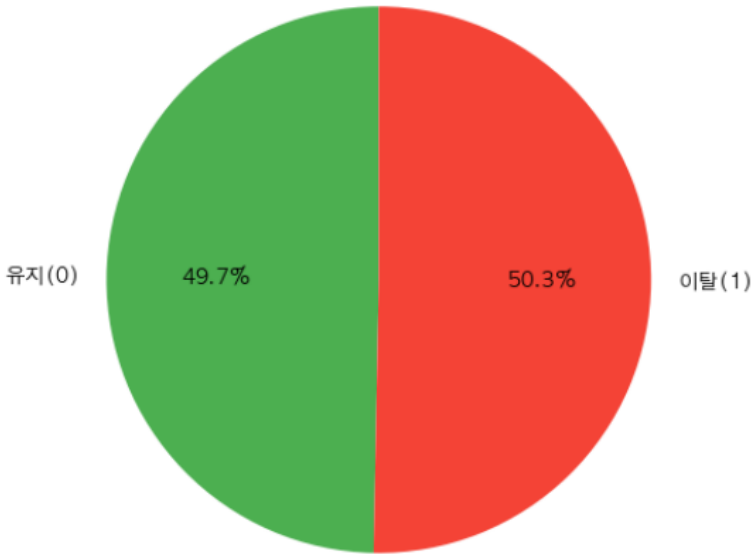
2. 탐색적 데이터 분석 (EDA)

- **기초 통계 확인:** 각 변수의 평균, 표준편차, 결측 비율 확인

- **데이터 분포 시각화**
  - **전체 고객 이탈 비율 분석 (churn)**

    전체 고객 중 **이탈률은 50.3%, 유지율은 49.7%**로 이탈 고객과 유지 고객의 비율이 거의 동일함.

    → 타깃 변수(churn)가 균형 잡힌 분포를 보이는 데이터셋으로 볼 수 있으며, 모델 학습 시 클래스 불균형에 따른 편향 우려가 낮음.

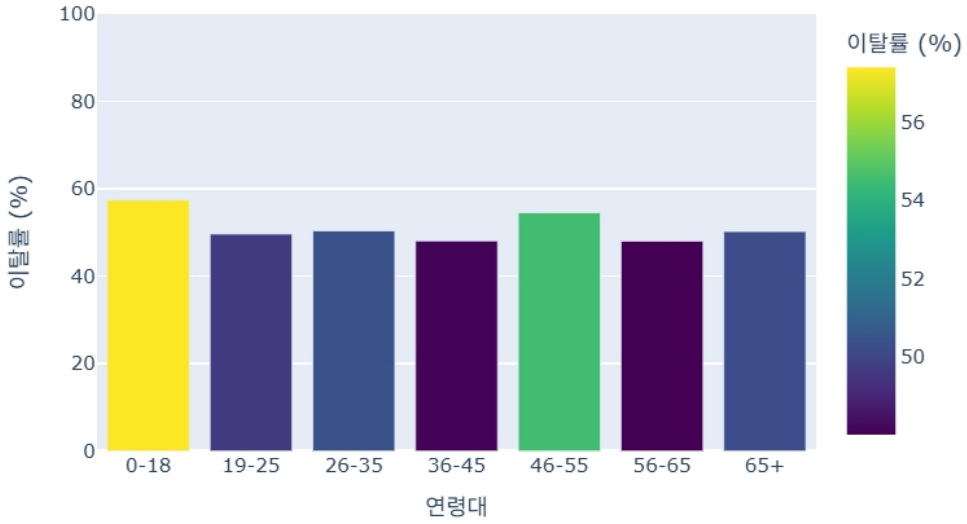


그래프1. 전체 고객 이탈 비율

- **연령대별 평균 이탈 비율 (age)**

  0-18세 구간이 가장 높고, 19-35세는 다소 낮으며 46세 이상부터 다시 약간 증가하는 U자형 패턴을 보임

  → 이는 젊은층의 결제 지속성이 낮고, 중장년층의 콘텐츠 피로도 감소 등 요인으로 해석 가능하며, 전반적으로 연령대가 비교적 균등하게 분포된 데이터셋으로 볼 수 있음.

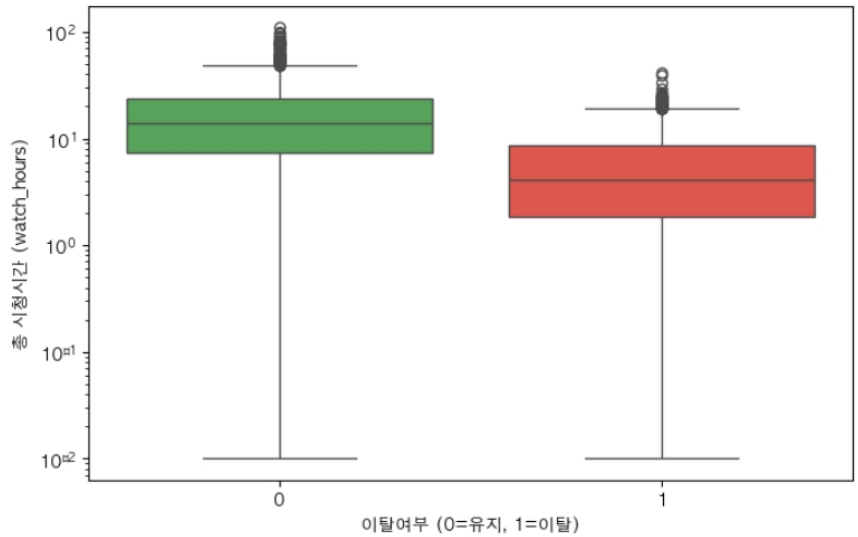


그래프2. 연령대별 평균 이탈률

- **시청 시간과 이탈 관계 (watch_hours)**

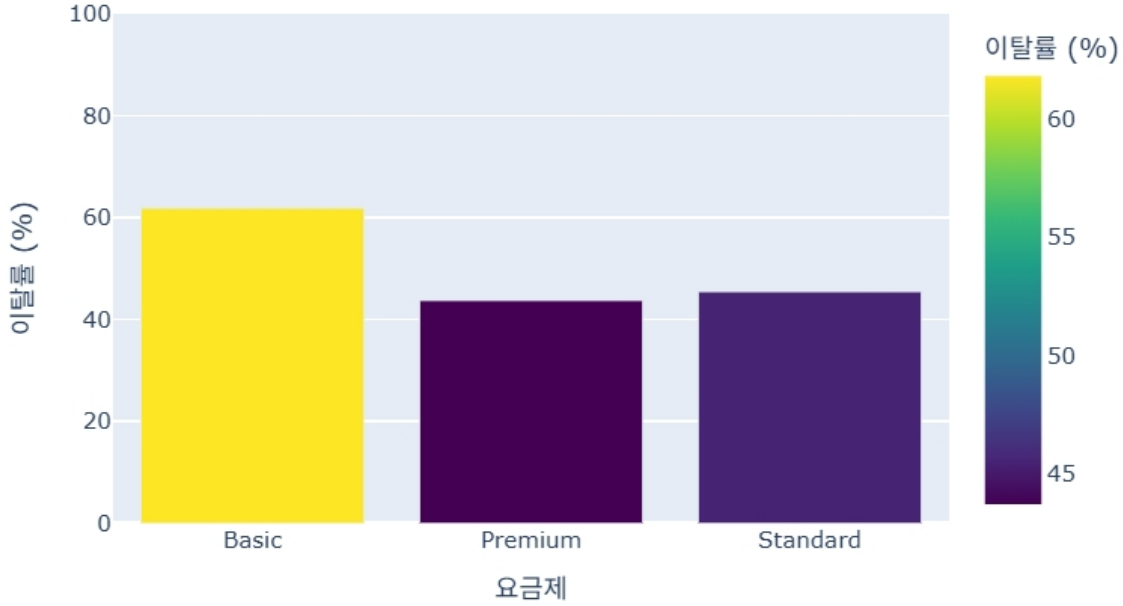
  왼쪽(초록색, 유지 고객)의 중앙값이 오른쪽(빨간색, 이탈 고객)보다 확연히 높게 나타남.

  → 시청 시간이 짧을수록 고객 이탈 가능성이 높다고 해석할 수 있음.



그래프3. 이탈 여부에 따른 시청시간 분포

- **요금제별 평균 이탈률 분석 (subscription_type)**
  Basic 요금제의 이탈률이 약 60%로 가장 높으며, Standard 및 Premium 요금제는 각각 약 45% 수준으로 상대적으로 낮게 나타남.
  → 저가 요금제(Basic) 고객의 충성도가 낮고 서비스 이탈 가능성이 높은 반면, Premium 요금제 고객은 만족도가 높거나 지속 이용 의사가 강한 경향으로 해석할 수 있음.



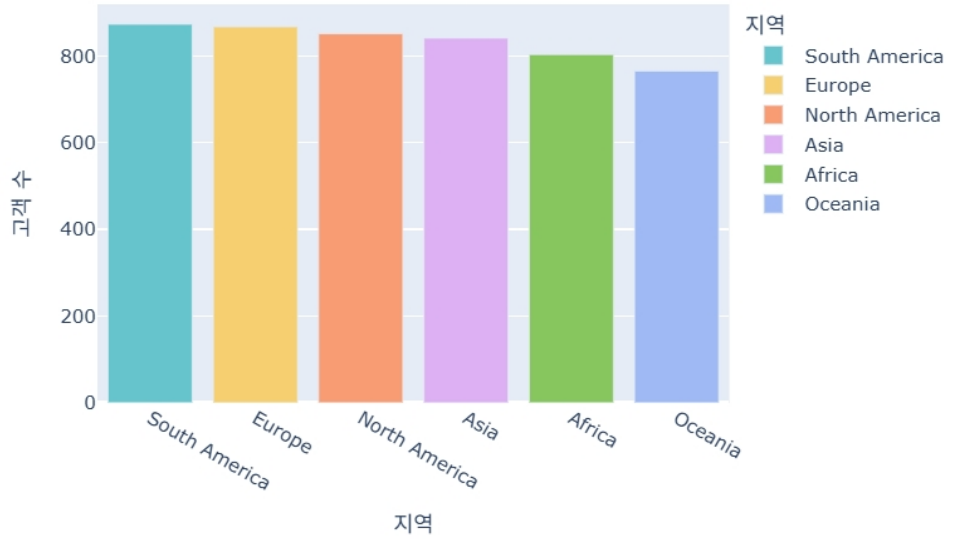
그래프4. 요금제별 평균 이탈률

- **지역별 고객 분포 및 이탈률 분석 (region)**

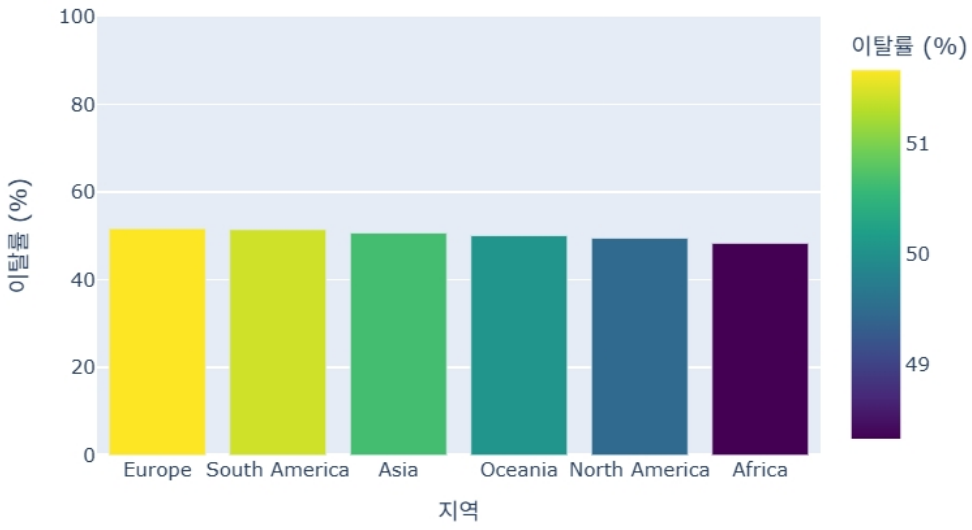
  South America, Europe, North America, Asia 순으로 고객 수가 많고, Oceania는 상대적으로 적게 분포함.

  전체적으로 지역 간 샘플이 균등하게 분포되어 있으며, 지역별 이탈률도 모두 약 50% 내외로 유사한 수준을 보임.

  → 이탈 요인이 지역적·문화적 요인보다는 요금제, 이용 행태, 시청 시간 등 개인적 요인에 더 큰 영향을 받는 데이터 특성으로 해석할 수 있음



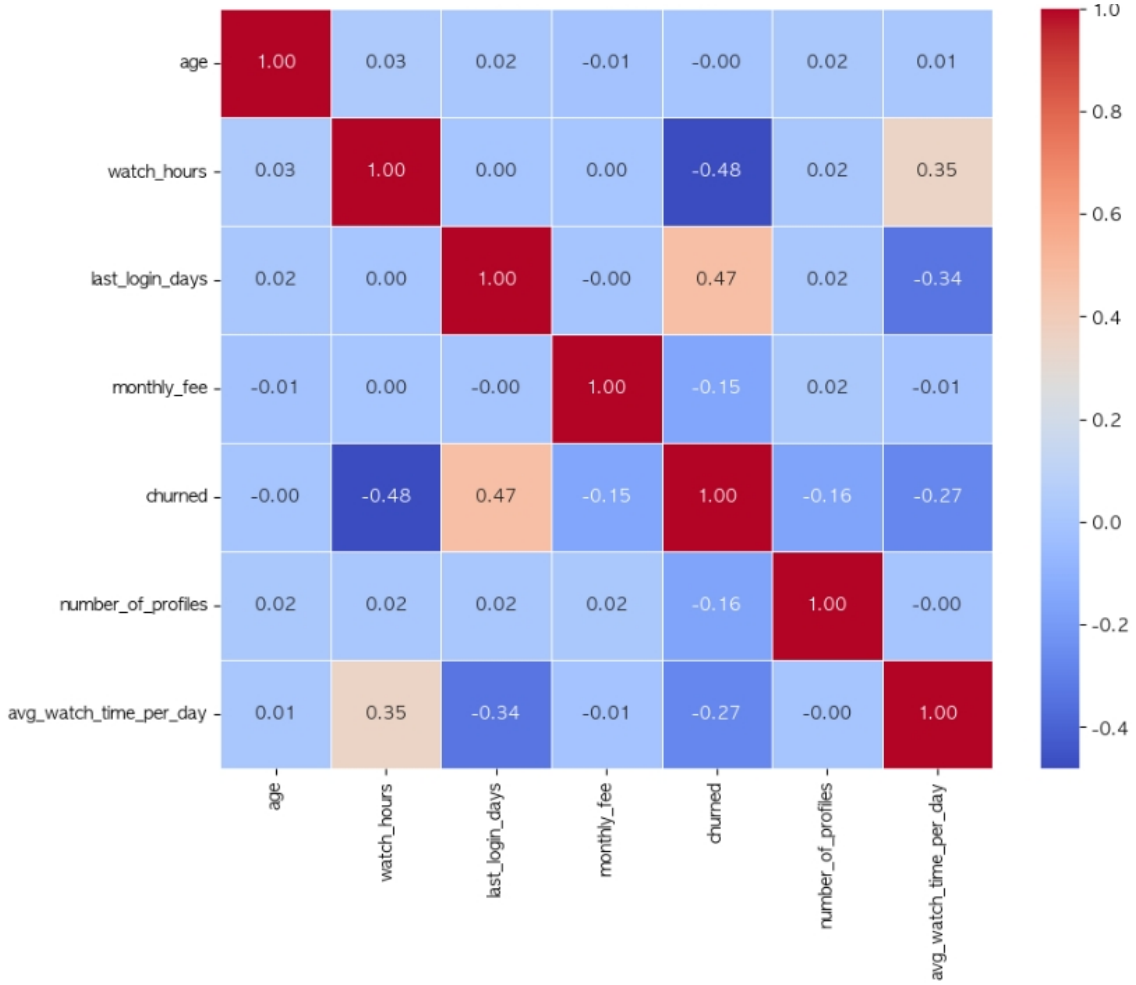
그래프5. 지역별 고객 수



그래프6. 지역별 평균 이탈률

● 상관관계 분석

- **heatmap을 통한 변수 간 상관관계 분석:**
  이용 빈도와 시청량이 감소하고 로그인 간격이 길어질수록 이탈 가능성이 급격히 증가하는 경향을 확인함.
  → 고객 행동 변화가 이탈에 직접적인 영향을 미치는 주요 요인임을 의미하며, 이러한 변수들은 이탈 조기 탐지 모델(Churn Prediction Model)의 핵심 feature로 활용 가능함.



그래프7. 수치형 변수 간 상관관계 히트맵

Ⅱ. 결측치 및 이상치 확인 및 처리

1. 결측치 확인 및 처리방법

- **결측치 확인 (df.isna().sum()) : 결측치 없음**

customer_id	0
age	0
gender	0
subscription_type	0
watch_hours	0
last_login_days	0
region	0
device	0
monthly_fee	0
churned	0
payment_method	0
number_of_profiles	0
avg_watch_time_per_day	0
favorite_genre	0

2. 이상치 확인 및 처리 방법

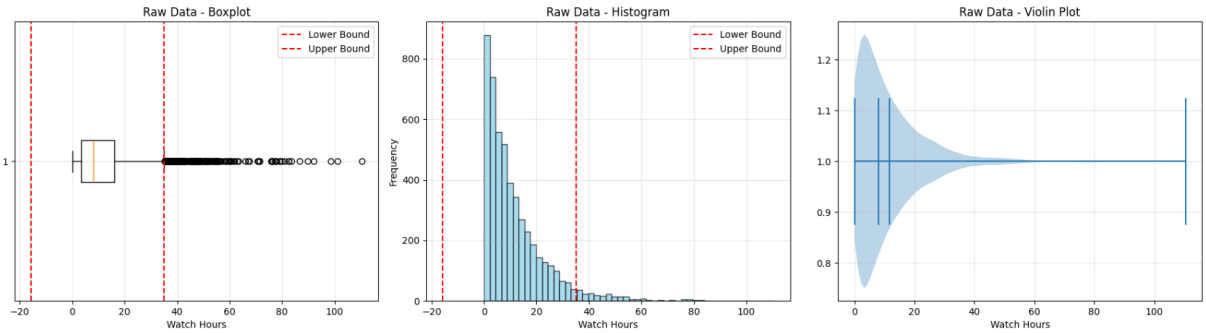
2-1. watch_hours (상관관계 -0.48)

- **이상치 확인 (IQR, Z-score)**

① 이상치 경계값:

- Lower Bound: -15.70
- Upper Bound: 35.07

② 이상치 개수: 238개 (4.76%)



그래프7. 이상치 확인 시각화(watch_hours)

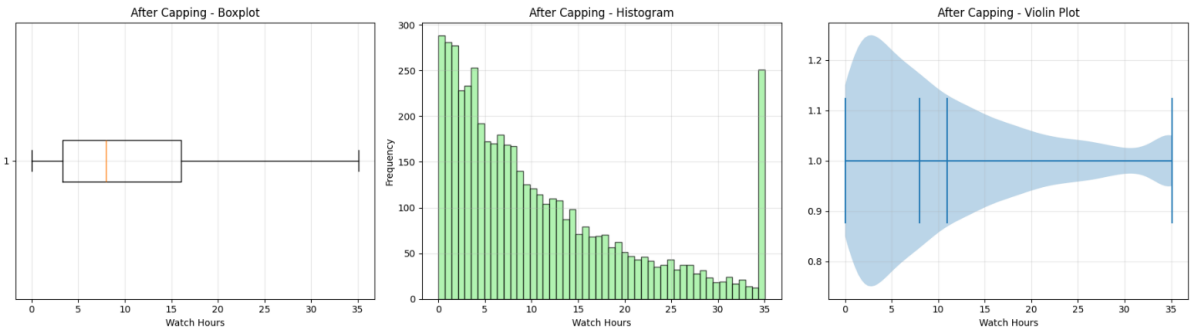
● **이상치 처리 (Capping 대체)**

① capping 처리

- -15.70보다 작은 값 → -15.70로 대체 (실제 0이하값 없음)
- 35.07보다 큰 값 → 35.07로 대체

② 통계량 비교:

	원본	처리 후
평균 (mean)	11.65	10.98
중앙값 (median)	8.00	8.00
표준편차 (std)	12.01	9.68
최소값 (min)	0.01	0.01
최대값 (max)	110.40	35.07



그래프8. 이상치 처리 후 데이터 시각화(watch_hours)

- **이상치 처리 결과 요약**
  - 이상치 238개(4.76%)를 Capping 처리하여 분포 안정화 및 왜곡 완화

2-2. last_login_days (상관관계 0.47)

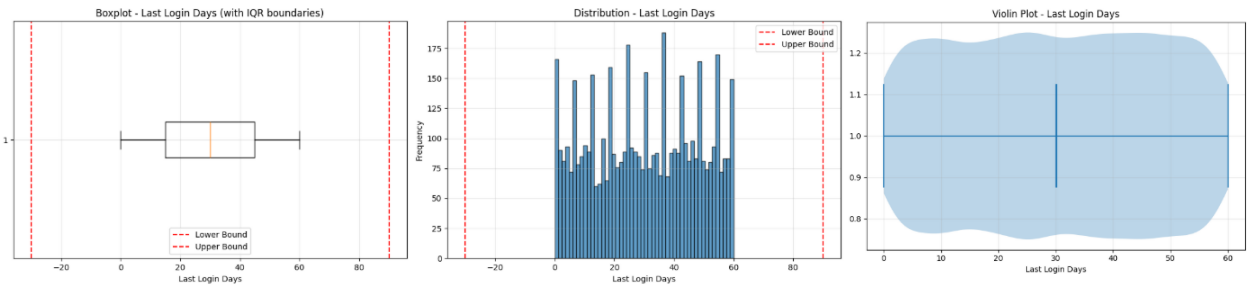
- **이상치 확인 (IQR, Z-score)**

  ① IQR 기반 이상치 경계:

  - Q1 (25%): 15.00
  - Q3 (75%): 45.00
  - IQR: 30.00
  - Lower Bound (Q1 - 1.5*IQR): -30.00
  - Upper Bound (Q3 + 1.5*IQR): 90.00

  ② IQR 기반 이상치: 0개 (0.00%)

  ③ Z-Score 기반 이상치 (|Z| > 3): 0개 (0.00%)



그래프9. 이상치 시각화(last_login_days)

- **이상치 확인 결과**
  - IQR 및 Z-Score 를 통해 이상치가 없음을 확인. 이상치 처리 사항 없음.

3. 결측치 및 이상치 처리 결과

- 결측치는 존재하지 않음
- watch_hours 변수의 이상치를 Capping 처리하여 분포 왜곡 완화
- last_login_days 이상치 없음
- 이상치 처리 결과, 데이터 분포의 안정성이 향상되어 모델 학습 시 신뢰도 개선 기대

Ⅲ. 데이터 정제 및 변환

1. 변수 정제 및 변환

- **불필요 컬럼 제거**

  ○ 제거 대상

  customer_id, monthly_fee, payment_method, avg_watch_time_per_day

  ○ 제거 사유

  - customer_id : 단순 식별자, 모델 학습에 불필요

  - monthly_fee : 요금정보가 subcription_type(구독타입)에 내재되어 있어 동일한 의미를 중복 반영하므로 제거

  - payment_method : 결제 수단으로서 예측 기여도 낮음

  - avg_watch_time_per_day : watch_hours(시청시간)와 유사하고 단위(time)의 기준이 불명확하여 제거

  ○ 불필요 컬럼 제거를 통해 데이터 차원을 축소하고 모델 효율성을 향상시킴.

- **변수 정제 및 변환**

① 범주형 변수 : One-hot Encoding 및 Label Encoding 적용

○ 적용 변수: gender, subscription_type, region, device, favorite_genre

○ 처리 방식: 문자열로 된 범주형 변수를 정수형으로 변환하여 모델이 학습할 수 있도록 함.

○ 세부 내용:

- gender(성별) : 'Female', 'Male', 'Other'
- subscription_type(구독 타입) : 'Basic', 'Standard', 'Premium'
- region(지역) : 대륙 단위로 구분
- device(시청 기기) : 'TV', 'Mobile', 'Laptop', 'Desktop', 'Tablet'
- favorite_genre(선호 장르) : 'Action', 'Sci-Fi', 'Drama', 'Horror', 'Romance', 'Comedy', 'Documentary'

○ 적용 사유: 모델이 범주형 데이터를 수치적으로 인식할 수 있도록 하되, 범주 간 서열 관계를 부여하지 않기 위함.

② 수치형 변수 : Scaling 적용

○ 적용 변수: watch_hours

○ 처리 방식: 평균 0, 표준편차 1로 표준화하여 모델 학습 시 안정적 수렴을 유도함.

○ 적용 사유: 변수 간 단위 차이를 제거하여 모델 성능을 향상시키기 위함

③ age 변수 : One-hot Encoding 및 Min-Max Scaling 적용

○ One-Hot Encoding 적용

■ 처리 방식: 연령을 구간화 후, 각 연령대를 독립된 범주로 변환.

■ 적용 사유: 자 크기에 따른 불필요한 가중치 부여를 방지하고, 연령대별 특성을 명확히 반영하기 위함.

○ Min-Max Scaling 적용

■ 처리 방식: 연속값을 0~1 범위로 정규화함.

■ 적용 사유: 다른 수치형 변수와 동일한 스케일로 맞춰 학습 안정성을 확보하기 위함.

3. 변환 결과

● 전처리 완료 I (/netflix_customer_onehot_preprocessed.csv)

- 범주형 One-Hot → watch_hours Standard Scaling → age One-Hot, Min-Max
- 최초 14개 컬럼 → 변환 후 38개 컬럼

● 전처리 완료 II (netflix_customer_churn_tree_preprocessed.csv)

- 범주형 Label
- 최초 14개 컬럼 → 변환 후 10개 컬럼

Ⅳ. 전처리 과정의 효율성 및 설명

- 전처리 과정에서 pandas (get_dummies) 와 scikit-learn (LabelEncoder) 등 표준 라이브러리를 활용하여 데이터 변환 및 정제 방법을 명확하고 효율적으로 수행했음.
- 단계별 로그 출력과 데이터 shape 확인으로 처리 과정의 추적성과 안정성을 높였음.
- EDA → 결측치 확인 → 이상치 처리 → 인코딩 → 스케일링의 순서로 일관되게 진행하였으며, 각 단계별 처리 근거와 결과를 수치로 제시하여 전처리 과정의 명확성과 타당성을 확보했음.
- 이상치 보정 및 불필요 변수 제거를 통해 데이터 품질을 개선하였고, 스케일링 및 인코딩을 통해 모델 입력값의 일관성을 확보했음.
- 결과적으로 데이터의 신뢰도를 향상시켜, 머신러닝 모델 및 딥러닝 모델에서 안정적인 학습과 성능 향상이 기대됨.

Ⅴ. 결론 및 요약

- **EDA 결과:** 주요 변수 간 상관관계 및 데이터 특성 파악 완료
- **결측/이상치 처리:** 결측치 확인 / 이상치 대체 및 유지 결정
- **정제/변환:** 인코딩·스케일링 적용으로 데이터 정규화
- **효율성:** pandas·sklearn 기반 함수화로 전처리 자동화 및 일관성 확보