**전처리 결과서**

1. **Explore Data**

텍스트, 폰트, 메뉴, 스크린샷이(가) 표시된 사진

자동 생성된 설명

* 총 고객 수(Rows) : 125,000
* 특성 수 (Columns) : 20

1. **결측치 확인**

텍스트, 스크린샷, 폰트, 문서이(가) 표시된 사진

자동 생성된 설명

* 결측치 확인 결과 : 0개

1. **IQR을 이용한 이상치 확인**

텍스트, 스크린샷, 폰트, 문서이(가) 표시된 사진

자동 생성된 설명

* 이상치 확인 결과 : 0개

1. **Feature Engineering**
   1. 평균 재생 시간 당 곡수(songs\_per\_hours)

: weekly\_songs\_played / weekly\_hours 로 파생변수 생성함.

* 1. 가입 이후 경과 일수(signup\_date\_positive)

: 현재 날짜 기준으로 언제 가입했는지에 대한 차이가 음수로 표현되어있었음. 따라서, 이를 양수로 전환하여 직관적으로 표현함.

* 1. 서비스 내 활동 지표(engagement\_score)

: 서비스 내에서 사용자의 활동을 확인할 수 있는 특성들인 notifications\_clicked와 num\_playlists\_created, num\_shared\_playlists 특성들을 평균 계산하여 파생변수 생성하였음.

1. **상관관계 시각화 : 상관관계를 한눈에 확인할 수 있는 heatmap을 그려보았음.**

텍스트, 스크린샷, 도표, 평행이(가) 표시된 사진

자동 생성된 설명

* 전반적으로 상관관계가 낮아보임.
* ‘churned’와 상대적으로 높은 상관관계를 보이는 특성은 ‘weekly\_hours’로 확인됨.

1. **원핫 인코딩**

: EDA로 파악한 결과, ‘churned’에 유의미하다고 판단된 범주형 특성들을 원핫 인코딩하였다.

* ‘subscription\_type’, ‘num\_subscription\_pauses’, ‘customer\_service\_inquiries’

1. **로그 변환 + 정규화**

: 많은 머신러닝 알고리즘은 데이터가 정규 분포를 따를 때 더 잘 작동함. 로그변환은 데이터의 분포를 정규 분포에 더 가깝게 만들어 줌. 우리가 확보한 데이터의 특성들의 크기가 다양하여서 로그 변환을 통해 특정 특성이 다른 특성보다 더 큰 영향을 미치지 않도록 함. 로그 변환 이후 정규화도 수행함.

* ‘age’, ‘signup\_date\_positive’, ‘weekly\_hours’, ‘average\_session\_length’, ‘weekly\_songs\_played’, ‘weekly\_unique\_songs’, ‘num\_favorite\_artists’, ‘num\_platform\_friends’, ‘num\_playlists\_created’, ‘num\_shared\_playlists’, ‘notifications\_clicked’, ‘songs\_per\_hours’, ‘engagement\_score’, ‘song\_skip\_rate’

1. **EDA(가설 설정)**
   * Q1. 젊은 층이 이탈할 가능성이 낮을까?

텍스트, 스크린샷, 직사각형, 그래프이(가) 표시된 사진

자동 생성된 설명

* 해석 : 예상과는 달리 젊은 층에 속하는 20대의 이탈률이 높았음. 그리고 예상대로 60, 70대의 이탈률 또한 높았음.
  + Q2. 무료 구독 사용자들이 유료 구독 사용자들보다 이탈할 가능성이 높을까?

텍스트, 스크린샷, 직사각형, 도표이(가) 표시된 사진

자동 생성된 설명

* 해석 : 예상대로 무료 구독 사용자들의 이탈률이 높았음(38.7%).
  + Q3. 특정 결제 방식을 사용하는 고객에 따라 이탈할 가능성이 다를까?

텍스트, 스크린샷, 직사각형, 그래프이(가) 표시된 사진

자동 생성된 설명

* 해석 : 결제 방식에 따른 이탈률의 차이가 미미함.
  + Q4. 고객 서비스 문의 빈도가 높은 고객들이 이탈할 가능성이 높을까?

스크린샷, 텍스트, 직사각형, 도표이(가) 표시된 사진

자동 생성된 설명

* 해석 : 고객 서비스 문의 빈도가 높을수록 이탈률이 높았음.
  + Q5. 구독 기간이 긴 고객들이 이탈할 가능성이 낮을까?

텍스트, 스크린샷, 직사각형, 그래프이(가) 표시된 사진

자동 생성된 설명

* 해석 : 구독 기간에 따른 이탈률의 차이가 미미함
  + Q6. 서비스 내에서 많은 친구를 가질수록 고객들이 이탈할 가능성이 낮을까?

텍스트, 스크린샷, 직사각형, 사각형이(가) 표시된 사진

자동 생성된 설명

* 해석 : 서비스 내에서의 친구 수에 따른 이탈률의 차이가 미미함.
  + Q7. 구독을 일시 중지한 횟수가 높을수록 이탈할 가능성이 높을까?

텍스트, 스크린샷, 직사각형, 도표이(가) 표시된 사진

자동 생성된 설명

* 해석 : 구독을 일시 중지한 횟수가 높을수록 이탈률이 높았음.
  + Q8. 1주당 서비스 이용시간이 많을수록 이탈할 가능성이 낮을까?

텍스트, 스크린샷, 직사각형, 도표이(가) 표시된 사진

자동 생성된 설명

* 해석 : 1주당 서비스 이용시간이 많을수록 이탈률이 낮았음.
  + Q9. 음악을 듣는 도중 스킵한 비율이 높을수록 이탈할 가능성이 높을까?

텍스트, 스크린샷, 직사각형, 도표이(가) 표시된 사진

자동 생성된 설명

* 해석 : 음악을 듣는 도중 스킵한 비율이 높을수록 이탈률이 높았음.

**9. 클러스터링**

* 70% 이상 설명할 수 있는 주성분 개수를 설정하기 위해 n\_components는 0.7로 설정하였음.
* 0과 1 이진 분류 문제이기 때문에, n\_clusters는 2로 설정하였음.
* PCA분석과 클러스터링 후 시각화 한 결과.

텍스트, 도표, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명