데이터전처리 결과서

팀명: 3팀

팀원: 강민지, 김지은, 김효빈, 안채연, 홍혜원

목차

데이터전처리 결과서 1

1. 기존 데이터셋 2

1.1. 컬럼 설명 2

1.2. 주요 컬럼 기술통계량 확인 2

1.3. 결측치 확인 3

2. 데이터 전처리 개요 3

2.1. 전처리 방향 3

2.2. 트리 기반 vs. 비트리 기반 모델의 데이터 처리 전략 4

3. EDA/트리 기반 모델용 데이터셋 4

3.1. 전처리 과정 4

3.1.1. 불필요한 행 제거 / 이상치 및 결측치 처리 4

3.1.2. 파생 변수 생성 4

3.2. 최종 데이터셋 6

4. 회귀 및 비트리 기반 모델용 데이터셋 6

4.1. 전처리 과정 6

4.1.1. 불필요한 행 제거 / 이상치 및 결측치 처리 6

4.1.2. 범주형 변수 원-핫 인코딩(One-Hot Encoding) 6

4.1.3. 연속형 변수 유지(Continuous Variables) 7

4.1.4. 로그 변환 (bill_avg, download_avg_log, upload_avg_log) 7

4.2. 최종 데이터셋 8

1. 기존 데이터셋

1.1.컬럼 설명

컬럼명	설명
`id`	userID
`is_tv_subscriber`	TV 구독 여부 [1:구독, 0:미구독]
`is_movie_package_subscriber`	영화 패키지(시네마) 구독 여부 [1:구독, 0:미구독]
`subscription_age`	서비스 이용 기간(년 단위), 고객이 서비스를 이용한 총 연수
`bill_avg`	최근 3개월 평균 청구 금액 (단위: $)
`remaining_contract`	남은 계약 기간(년 단위), **null이면 계약 없음**. 계약 중 고객은 중도 해지 시 위약금 발생
`service_failure_count`	최근 3개월 동안 고객센터에 신고한 서비스 장애 건수 / 서비스 품질 관련 지표
`download_avg`	최근 3개월 평균 다운로드 사용량(GB) / 인터넷 사용량 지표
`upload_avg`	최근 3개월 평균 업로드 사용량(GB) / 인터넷 업로드 사용량
`download_over_limit`	지난 9개월 동안 다운로드 한도 초과 횟수, 한도 초과 시 추가 요금 발생
`churn`	이탈 여부(**target**) [1:서비스 해지, 0:서비스 유지]

- 데이터 형태: (72274, 11)

1.2.주요 컬럼 기술통계량 확인

	is_tv_subscriber	is_movie_package_subscriber	subscription_age	reamining_contract	service_failure_count
count	72274.000000	72274.000000	72274.000000	50702.000000	72274.000000
mean	0.815259	0.334629	2.450051	0.716039	0.274234
std	0.388090	0.471864	2.034990	0.697102	0.816621
min	0.000000	0.000000	-0.020000	0.000000	0.000000
25%	1.000000	0.000000	0.930000	0.000000	0.000000
50%	1.000000	0.000000	1.980000	0.570000	0.000000
75%	1.000000	1.000000	3.300000	1.310000	0.000000
max	1.000000	1.000000	12.800000	2.920000	19.000000

download_avg	upload_avg	download_over_limit
71893.000000	71893.000000	72274.000000
43.689911	4.192076	0.207613
63.405963	9.818896	0.997123
0.000000	0.000000	0.000000
6.700000	0.500000	0.000000
27.800000	2.100000	0.000000
60.500000	4.800000	0.000000
4415.200000	453.300000	7.000000

1.3. 결측치 확인

	0
id	0
is_tv_subscriber	0
is_movie_package_subscriber	0
subscription_age	0
bill_avg	0
reamining_contract	21572
service_failure_count	0
download_avg	381
upload_avg	381
download_over_limit	0
churn	0

dtype: int64

- remaining_contract: null 은 '계약 없음'을 뜻한다.
- download_avg 의 결측치 행과 upload_avg 의 결측치 행이 동일한 것을 확인하였다.

2. 데이터 전처리 개요

2.1. 전처리 방향

본 프로젝트의 전처리 단계는 이탈(churn) 예측 모델의 성능 향상과 데이터 해석력 강화를 목표로 수행하였다. 원본 데이터는 고객의 구독 상태, 사용 행태, 계약 정보 등 다양한 유형의 변수를 포함하고 있었으며, 이를 모델 특성에 맞게 구조화하고 이상치를 정제하는 과정이 필요했다.

특히 본 데이터셋은 연속형 변수와 범주형 변수가 혼합되어 있었기 때문에, 모델 유형에 따라 인코딩 및 변환 전략을 구분하여 적용하였다. 전처리 방향은 다음과 같다.

- 데이터 품질 확보: 결측치, 이상치, 불필요 컬럼 제거
- 범주형 변수 처리 최적화: 모델 유형(Tree / Non-Tree)에 따른 인코딩 방식 분리
- 수치형 변수 분포 안정화: 로그 변환을 통한 왜도(skewness) 완화

2.2.트리 기반 vs. 비트리 기반 모델의 데이터 처리 전략

구분	트리 기반 모델 (Tree-based Models)	비트리 기반 모델 (Non-tree Models)
주요 모델	Decision Tree, Random Forest, XGBoost	Logistic Regression, SVM
데이터셋 명	`df_tree`	`df_re`, `df_re_log`
범주형 변수 처리	**Label Encoding** (범주를 순서형 정수로 변환)	**One-hot Encoding** (더미 변수 생성)
연속형 변수	구간화(binning) 적용 (`subscription_age_group`)	원본 값 유지 (`subscription_age`)
로그 변환	불필요 (트리모델은 분포 형태 영향 적음)	`bill_avg`, `download_avg`, `upload_avg` 에 적용
목적	규칙 기반 의사결정 구조 파악	연속적 확률 예측 및 선형 분리 성능 향상

3. EDA/트리 기반 모델용 데이터셋

트리 기반 모델(Decision Tree, Random Forest, XGBoost 등)은 데이터의 비선형 구조를 자동으로 학습하기 때문에, 범주형 변수의 순서형 인코딩(Label Encoding) 과 연속형 변수의 구간화(Binning) 만으로도 충분히 의미 있는 패턴을 학습할 수 있다.
이에 따라 본 데이터셋은 EDA(탐색적 데이터 분석) 과 트리 기반 모델 학습을 위해 다음과 같은 전처리를 수행하였다.

3.1.전처리 과정

3.1.1. 불필요한 행 제거 / 이상치 및 결측치 처리

- id 컬럼은 모델 학습에 불필요하므로 삭제하였다.
- dwonload_avg 가 NaN 인 행과 upload_avg 가 NaN 인 행이 일치함을 확인한 후, 해당 행 381 개를 제거하였다.
- subscription_age 의 최솟값이 -0.02 이므로, 해당 행을 제거하였다.
- bill_avg, download_avg, upload_avg 의 분포를 확인한 결과, 극단값이 존재했으나 트리 모델은 이상치에 덜 민감하므로 제거하지 않았다.
- 대신 통계 검증 및 로그 변환 분석용 데이터셋(df_re_log)에서 별도로 비교하였다.

3.1.2. 파생 변수 생성

3.1.2.1. 계약 상태 라벨링 (contract_type)

원본 컬럼 remaining_contract 는 잔여 계약 기간을 수치형으로 제공하지만, 결측치(NaN)가 많고 최소값 0 과 양수 값이 혼재되어 있었다.

- 이 정보를 계약 상태로 단순화하여 다음 세 가지 범주로 재구성하였다.

조건	새로운 라벨	의미
`remaining_contract` is `null`	`no_contract`	무약정 (자유이용 고객)
`remaining_contract == 0`	`expired`	계약 종료 고객
`remaining_contract > 0`	`active`	약정 유지 중인 고객

- 이후 모델 학습을 위해 정수형으로 라벨링하였다. {'no_contract': 0, 'expired': 1, 'active': 2}
- 기존 remaining_contract 컬럼을 삭제하였다.

3.1.2.2. 구독 연수 구간화 (subscription_age_group)

- subscription_age 는 고객이 서비스를 이용한 총 연수로, 연속형 변수이지만 고객 충성도와 약정 주기를 반영하기 위해 4 단계 구간화(binning) 를 적용하였다.

구간	라벨	고객군 정의
$0 \le x < 1$	`0~1년`	신규 또는 초기 고객
$1 \le x \le 3$	`1~3년`	약정 중기 고객
$3 < x \le 5$	`3~5년`	재계약 고객
$x > 5$	`5년 초과`	충성 고객

- 이후 모델 학습을 위해 정수형으로 라벨링하였다. {'03 년': 1, '3~5 년': 2, '5 년 이상': 3}
- 기존 subscription_age 컬럼을 삭제하였다

3.1.2.3. 구독 유형 통합 (subscription_label)

- is_tv_subscriber 와 is_movie_package_subscriber 는 각각 tv 구독과 영화패키지 구독이다 두 서비스의 조합에 따라 고객 행동 패턴이 달라질 가능성을 알아보기 위해 두 변수를 통합하여 하나의 구독 조합 컬럼을 생성하였다.

TV 구독	영화 구독	통합 라벨 (`subscription_label`)	의미
0	0	`none`	구독 없음
1	0	`tv`	TV만 구독
0	1	`movie`	영화만 구독
1	1	`both`	둘 다 구독

- 이후 모델 학습을 위해 정수형으로 라벨링하였다. {'none': 0, 'tv': 1, 'movie': 2, 'both': 3}
- 기존 is_tv_subscriber, is_movie_package_subscriber 컬럼을 삭제하였다.

3.2. 최종 데이터셋

항목	값
데이터 크기	(71892, 9)
주요 컬럼	`bill_avg` , `service_failure_count` , `download_avg` , `upload_avg` , `download_over_limit` , `churn` , `contract_type` , `subscription_age_group` , `subscription_label`
타깃 컬럼	`churn` (1=이탈,0=유지)

4. 회귀 및 비트리 기반 모델용 데이터셋

회귀 및 비트리 기반 모델(Logistic Regression, SVM 등)은 데이터의 선형 관계와 분포 형태에 민감하기 때문에, 트리 기반 모델(df_tree)과는 달리 보다 정규화한 입력 형태로 전처리를 진행하였다.

이 과정에서는 연속형 변수 유지, 원-핫 인코딩(One-Hot Encoding), 로그 변환(Log Transformation)을 중심으로 데이터셋을 구성하였다.

4.1. 전처리 과정

4.1.1. 불필요한 행 제거 / 이상치 및 결측치 처리

df_tree 와 동일하다.

4.1.2. 범주형 변수 원-핫 인코딩(One-Hot Encoding)

- 회귀 및 비트리 기반 모델에서는 범주의 순서가 존재하지 않으므로, 더미 변수(0/1) 기반 인코딩으로 변환하였다.

원본 컬럼	인코딩 후 컬럼	처리 목적
`contract_type`	`contract_no_contract` , `contract_expired` , `contract_active`	계약 상태별 독립 변수 생성
`subscription_label`	`sub_none` , `sub_tv` , `sub_movie` , `sub_both`	구독 조합별 영향 학습 가능

- 이후 모델 학습을 위해 모든 더미 컬럼을 int 타입으로 변환하였다.
- 기존 contract_type, subscription_label 컬럼은 삭제하였다.

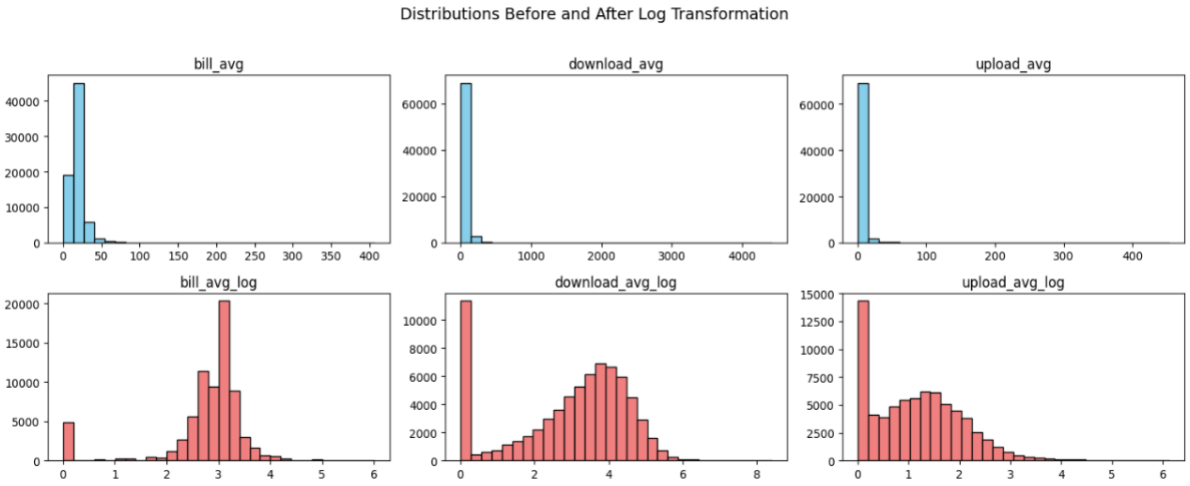
4.1.3. 연속형 변수 유지(Continuous Variables)

- 회귀 및 비트리 모델에서는 시간적 연속성을 반영하기 위해, 구간화하지 않고 원본 값 그대로 유지하였다.

컬럼	설명	처리 방식
`subscription_age`	서비스 이용 기간(년 단위)	원본 값 유지

4.1.4. 로그 변환 (bill_avg, download_avg_log, upload_avg_log)

- bill_avg, download_avg, upload_avg 컬럼은 분포의 왜도(skewness)가 크고, 상위 이상치가 모델 학습에 영향을 줄 가능성이 있었다.
- 따라서 로그 변환을 적용한 별도 데이터셋(df_re_log)을 생성하였다.



4.2.최종 데이터셋

항목	`df_ml_raw`	`df_ml_log`
데이터 크기	(71892, 14)	(71892, 14)
주요 컬럼	`bill_avg`, `download_avg`, `upload_avg`, `subscription_age`, `service_failure_count`, `download_over_limit`, `churn`	동일
추가 더미 컬럼	`contract_no_contract`, `contract_expired`, `contract_active`, `sub_none`, `sub_tv`, `sub_movie`, `sub_both` (총 7개)	동일
로그 변환 컬럼	없음	`bill_avg_log`, `download_avg_log`, `upload_avg_log`
타깃 컬럼	`churn` (1=이탈, 0=유지)	동일