온라인 쇼핑 세션 구매 확률 예측 모델 – 데이터 전처리 보고서

데이터 전처리 보고서

(UCI Online Shoppers Purchasing Intention Dataset 기반)

데이터 전처리 목적

온라인 쇼핑 세션 데이터를 모델 학습이 가능한 형태로 가공하는 과정을 기록

의사결정 근거와 데이터 특성에 대한 해석을 명확히 설명하기 위함

- 원본 데이터가 어떤 특성을 갖고 있었는지
- 어떤 문제가 있었고 (결측치, 이상치, 불균형 등)
- 왜 특정 전처리 전략을 선택했는지

1. 데이터 이해 및 탐색 (EDA)

1.1 원본 데이터 개요

- 데이터셋: **UCI Online Shoppers Purchasing Intention Dataset**
- 관측 단위: **1행 = 1 세션(Session)**
- 데이터 성격:
  - 온라인 쇼핑 웹사이트 방문 로그 기반
  - 실제 이커머스 환경과 유사한 행동 데이터
- 타겟 변수: `Revenue`
  - `0`: 해당 세션에서 구매 발생하지 않음
  - `1`: 해당 세션에서 구매 발생

본 데이터는 실제 운영 환경과 마찬가지로 **구매가 발생한 세션의 비율이 낮은 불균형 데이터**라는 특징

1.2 컬럼 구성 및 정의

컬럼은 크게 세 가지 그룹으로 구성되어 있다.

① 행동 관련 수치형 변수

컬럼명	의미	타입	단위
Administrative	관리 페이지 방문 횟수	int	count
Informational	정보 페이지 방문 횟수	int	count
ProductRelated	상품 페이지 방문 횟수	int	count
Administrative_Duration	관리 페이지 체류 시간	float	seconds
Informational_Duration	정보 페이지 체류 시간	float	seconds
ProductRelated_Duration	상품 페이지 체류 시간	float	seconds

→ **어떤 유형의 페이지를 얼마나 자주, 얼마나 오래 봤는지**를 나타내는 핵심 행동 지표

---

② 세션 품질 관련 변수

컬럼명	의미	타입	단위
BounceRates	단일 페이지 방문 비율	float	ratio
ExitRates	종료 페이지 비율	float	ratio
PageValues	페이지의 기대 수익 기여도	float	value
SpecialDay	특수일(기념일) 근접도	float	ratio

---

③ 세션 / 방문자 속성 변수 (범주형)

컬럼명	의미	타입
Month	방문 월	category
OperatingSystems	OS 유형	category
Browser	브라우저 유형	category
Region	지역 코드	category
TrafficType	유입 채널 유형	category
VisitorType	방문자 유형 (New / Returning 등)	category
Weekend	주말 여부	bool

1.3 타겟 변수 정의

- `Revenue`
  - **이진 분류 문제**
  - 1: 구매 발생
  - 0: 구매 미발생

구매(1) 비율이 낮아 **Accuracy 중심의 평가나 전처리는 부적절**하며, 모델링 및 데이터 분할 전략에 직접적인 영향을 미친

1.4 기초 통계 및 데이터 특성

- 수치형 변수들은 **0 값이 많음**
  - 이는 "데이터 누락"이 아니라 **해당 행동이 발생하지 않은 세션**을 의미
- 체류 시간 관련 변수는 **우측 꼬리가 긴 분포(right-skewed)**를 보임
- PageValues는 구매와 직접적으로 연관된 변수로, 값이 0인 경우가 다수

1.5 결측치 분포

- 명시적인 NaN 결측치 없음
- 다만 수치형 변수에서의 `0` 값은 다음 두 가지 의미를 가짐
  - 실제 값 0
  - 행동 자체가 발생하지 않음

본 프로젝트에서는 이를 **구조적 결측(structural missing)** 으로 해석하였다.

1.6 이상치 확인

- 체류 시간(Duration) 계열 변수에서 일부 극단값 존재함
- 하지만:
  - 실제 사용자 행동에서 발생 가능한 값
  - 구매 전 장시간 탐색한 고객일 가능성 존재

→ **기계적 제거는 수행하지 않음**

1.7 클래스 불균형 여부

- 구매(Revenue=1) 비율이 전체의 소수

- 심각한 클래스 불균형 존재

이는 이후:

- 모델 선택
- 평가 지표(PR-AUC 중심)
- Stratified Split 사용

  의 핵심 근거가 된다.

---

2. 전처리 전략 및 의사결정 근거

2.1 결측치 처리 전략

- 결측치 없음으로 별도 전략 진행하지 않음

2.2 이상치 처리 전략

- 이상치 탐지 후 제거 ❌
- 이유:
  - 실제 세션 행동의 다양성을 반영
  - 구매 가능성이 높은 장시간 탐색 세션 제거 위험
- 대응 방식:
  - **RobustScaler 사용**으로 스케일링 단계에서 영향 완화

2.3 범주형 변수 처리

- 선택 방식: **One-Hot Encoding**
- 사용 도구: `OneHotEncoder(handle_unknown="ignore")`

선택 이유

- 트리 기반 모델과의 궁합이 좋음
- Label Encoding은 순서 정보가 없는데도 순서를 부여할 위험 존재
- 새로운 카테고리 등장 시에도 에러 없이 처리 가능

2.4 수치형 변수 스케일링

- 적용 방식: **RobustScaler**
- 선택 이유:
  - 평균/표준편차 기반 스케일링보다 이상치에 강함
  - 체류 시간 등 극단값 영향을 완화하면서 정보는 보존

---

3. Feature Engineering

3.1 파생 변수 생성

- 본 프로젝트에서는 **과도한 파생 변수 생성은 지양**
- 이유:
  - 트리 기반 모델 특성상 비선형 패턴을 자체적으로 학습 가능
  - 해석 가능성과 과적합 리스크 관리

3.2 불필요한 컬럼 제거

- 명확한 식별자(ID) 컬럼 없음
- 타깃 누설(leakage) 위험 컬럼 없음

→ **원본 컬럼 대부분 유지**

3.3 Feature Selection 여부

- 명시적 Feature Selection 알고리즘 미적용
- 대신:
  - 트리 기반 모델의 내부 feature importance 활용
  - 모델 단계에서 자동으로 중요 변수 학습

---

4. 데이터 분할 전략

4.1 데이터 분할

모델 학습 및 평가를 위해 데이터를 다음과 같이 분할

- `train.csv` : base 모델 학습용
- `calib.csv` : 확률 보정(Calibration)용

- `test.csv` : 최종 성능 평가용

각 split은 타깃(Revenue) 비율이 유지되도록 stratified split으로 수행

4.2 Stratified Split 사용 여부

- **Stratified Split 사용**
- 기준 컬럼: `Revenue`

사용 이유

- 구매(1) 비율이 매우 낮은 불균형 데이터
- Train/Test 간 클래스 분포 불일치 시 성능 평가 왜곡 가능
- 각 데이터셋에서 **구매/미구매 비율을 동일하게 유지**하기 위함

요약

- 본 전처리 과정은 **데이터 의미를 훼손하지 않는 방향**을 최우선 기준으로 설계됨
- 이상치를 "무조건 제거"하지 않고, **비즈니스 및 도메인 관점에서 해석**
- 이후 모델링 단계(PR-AUC, F1 최적화)의 기반이 되는 **안정적이고 설명 가능한 데이터셋**을 생성함
