모델 정의서

XGBoost 기반 PPL 매출 예측 모델

1) 모델 개요

- **모델명**: XGBoost 기반 PPL 매출 예측 모델
- **목표**: 특정 산업군 내 PPL(Product Placement) 캠페인의 주요 특성(노출 장면 수, 노출 시간, 시청률, 사전 인식·구매 의도, 비용 등)을 기반으로 **매출 증가액을 예측**
- **배경**: 방송 및 OTT 콘텐츠 내 PPL의 효과성을 정량적으로 평가하고, 광고/마케팅 의사결정을 지원하기 위해 개발됨
- **범위**:
  - 데이터는 특정 산업군(예: 여행/레저, F&B 등)에 대해 필터링하여 학습 및 예측 수행
  - 최소 20건 이상의 샘플이 확보된 경우에만 모델 학습 가능
- **적용 분야**:
  - 광고·마케팅 기획 및 사전 ROI 예측
  - PPL 집행 후 효과 분석 및 리포팅
- **예상 결과**:
  - 총 노출 시간 변화에 따른 매출 증가액 추정 그래프 제공
  - 평균 ROI 및 산업군별 효과 요약 보고서 생성

2) 데이터 정보

- **출처**:
  - AWS RDS(MySQL) 내 `PPL_Service` 데이터베이스
  - 주요 테이블: `ppl_dummy_data`
- **형식**:
  - SQL DB → Pandas DataFrame 로드

- 정형 데이터 (행: 캠페인 단위, 열: 캠페인 특성/성과)

- **규모**:
  - 산업군별 최소 20건 이상 확보 필요
  - 예시 코드에서는 `"여행/레저"` 산업군 데이터 사용

특징 (주요 컬럼)

구분	컬럼명	설명
입력 변수 (Features)	`num_ppl_scenes`	PPL 장면 수
	`total_exposure_seconds`	총 노출 시간(초)
	`avg_viewer_rating_percent`	평균 시청률(%)
	`brand_awareness_pre_percent`	사전 브랜드 인식도(%)
	`purchase_intent_pre_percent`	사전 구매 의도(%)
	`production_cost_won`	제작비(원)
	`media_cost_won`	미디어 집행비(원)
타겟 변수 (Target)	`sales_increase_won`	매출 증가액(원)
평가용 지표	`roi_percent`	ROI(%)

- **전처리 과정**:
  - 결측치: `.fillna(0)` 처리
  - 범주형 변수 없음 (모두 수치형)
  - 산업군 필터링 (`WHERE industry = '{selected_industry}'`)

> **데이터 품질**:
> 더미 데이터 기반 (테이블명 ppl_dummy_data)
> 실제 산업 데이터 반영 시, 노이즈·편향 검증 필요

3) 모델 알고리즘

- **알고리즘 종류**:
  - `XGBRegressor` (XGBoost 회귀 모델)
- **주요 파라미터**:
  - `n_estimators = 500` (트리 개수)
  - `learning_rate = 0.05` (학습률)
  - `max_depth = 6` (트리 최대 깊이)
  - `random_state = 42` (재현성 확보)
- **학습 방식**:
  - 독립 변수: 캠페인 특성 데이터 (`features`)
  - 종속 변수: 매출 증가액(`sales_increase_won`)
  - 데이터셋 분할 없이 전체 데이터를 학습에 활용 (→ 추후 개선 필요: train/test split, k-fold cross validation)
- **평가 지표**:
  - 코드 상 RMSE/MAE 등은 직접 계산하지 않았음
  - 보고서에서는 **평균 ROI**와 **예측 매출 증가액**을 활용해 성과 해석

4) 모델 평가

- **성능 평가 지표 결과**:
  - 코드 상 성능 지표 출력 없음
  - 보고서 요약에는 다음 항목 포함
    - 평균 ROI (%)
    - 평균 예측 매출 증가액 (원)
    - 노출 시간 대비 매출 증가 예측 그래프
- **평가 방법**:
  - 현재 버전: 단순 학습 후 예측값 확인 (self-evaluation 수준)
  - 개선 필요: Hold-out 검증, 교차검증, baseline 모델 대비 성능 비교
- **결과 해석**:
  - 산업군별 평균 ROI와 예측 매출 증가액을 통해 캠페인 효과성을 직관적으로 파악 가능

- `total_exposure_seconds` (총 노출 시간) 변화에 따른 **what-if 분석**을 통해 PPL 기획 단계에서 효과 예측 가능
- 단, 과적합 여부 및 산업군별 데이터 편향은 향후 추가 검증 필요

> **향후 개선 사항**:
> • 데이터셋 분할을 통한 적절한 모델 검증
> • 다양한 성능 지표 도입 (RMSE, MAE, R^2)
> • 교차검증을 통한 모델 안정성 확보
> • 실제 데이터 적용 시 편향 및 노이즈 처리 방안 수립