**데이터 전처리 계획서**

작성일: 2025년 10월 13일  
작성자: 강성빈

# 1. 개요

본 문서는 삼성전자 IR 자료, DART 공시 데이터, 공공데이터포털 등 다양한 출처의 데이터를 통합하여 시계열 기반 예측 모델 학습에 적합한 데이터셋을 구축하기 위한 전처리 계획을 정의한다. 전처리 과정은 데이터 수집–정규화–결측치 처리–품질 검증–탐색적 분석–피처 엔지니어링 순으로 진행되며, 최종 산출물은 모델 학습용 Wide/Long 포맷 데이터셋이다.

# 2. 데이터 소스

데이터는 세 가지 주요 출처에서 가져왔습다.

|  |  |  |
| --- | --- | --- |
| 삼성전자IR PDF | 2020~2025년 분기별 컨퍼런스콜 자료 (매출, 영업이익, 부문별 실적, 재무상태표, 현금흐름표) | https://www.samsung.com/sec/ir/financial-information/earnings-release/ |
| 전자공시시스템 | 분기별 단일계정 재무제표 데이터 (한국채택국제회계기준, 원 단위 제공) | https://opendart.fss.or.kr/api/fnlttSinglAcntAll.json |
| 공공데이터포털 | 산업 동향, 출하 지수, 환율 등 외부지표 | https://api.odcloud.kr/api |

# 3. 데이터 적재 및 구조

데이터는 전처리 단계별로 구분된 폴더 구조를 통해 관리되며, 각 단계별 목적은 다음과 같다:

* Raw: 원본 PDF 텍스트 및 API 응답값을 포함한 비정형 데이터.
* Processed: 정규화 및 스키마 표준화(`period, metric, value`)가 완료된 중간 데이터.
* Final: 모델 학습에 최적화된 Wide 및 Long 포맷의 구조화 데이터.

# 4. 전처리 규칙

데이터의 일관성과 재현성을 확보하기 위해 다음과 같은 표준화 규칙을 적용하였다.

* 기간 표준화: ‘2024Q1’, ‘2025Q2’ 등으로 통일하여 정렬 및 병합 용이성을 확보.
* 단위 통일: 모든 금액은 ‘조원’ 단위로 변환 (DART 원 단위 → ÷1e12).
* 컬럼 구조 통일: IR, DART, ODcloud 데이터를 공통 스키마(`period, metric, value, category, unit`)로 변환.
* 세그먼트명 매핑: DX, MX/NW, DS, SDC, Harman 등 Canonical 라벨 기준으로 정규화.
* 결측치 처리: 원본 결측치는 유지하되, 모델 학습 시 lag/rolling feature로 보완.
* 중복 제거: 파일명, 페이지 정보 등 비식별 컬럼 제거 후 ‘period+metric’ 기준 중복 제거.
* 값 검증: 매출·영업이익 등 주요 지표는 정상 범위(매출 50~150조, 영업이익 0~40조) 내로 필터링.

# 5. 탐색적 데이터 분석 (EDA)

* 결측치 히트맵을 생성하여 변수별 누락 구간을 시각화.
* 기초 통계량(평균, 표준편차, 사분위수) 분석으로 데이터 스케일 확인.
* 피처 간 상관계수를 계산하여 다중공선성 여부를 판단.
* 시계열 그래프를 통해 매출·영업이익 추세를 분석하고 변동성을 파악.

# 6. 피처 엔지니어링

* Lag Feature: 이전 1~2개 분기의 주요 지표값 추가.
* Rolling Mean: 직전 2개 분기의 평균값을 계산하여 트렌드 완화.
* 외부지표 결합: 산업 출하지수 및 환율 데이터를 분기 단위로 집계 후 병합.
* 범주형 처리: 세그먼트명을 Canonical 라벨 기준으로 인코딩.

# 7. 품질 관리 및 로그

* PDF 파싱 시 각 페이지별 텍스트 덤프를 별도 debug디렉토리에 저장하여 추출 실패 구간을 추적.
* 현금흐름표 대사 검증: (기초현금 + CFO + CFI + CFF) ≈ 기말현금, 허용오차 ±0.1조.
* 결측치 및 이상치 검증 로그를 자동 저장(log.txt).

# 8. 산출물

* 전처리 완료 데이터셋: ml\_dataset\_wide.csv, ml\_dataset\_long.csv
* EDA 리포트: 결측치 현황, 상관분석, 시계열 그래프 포함.
* 모델 입력용 Feature 세트: Lag/Rolling 피처 반영 joblib 파일.

# 9. 기술 환경 및 스택

* 언어 및 버전: Python 3.13
* 사용 라이브러리: pandas, numpy, matplotlib, seaborn, pathlib, regex, joblib
* 개발 환경: Windows 11, VSCode
* 폴더 구조 자동화: Pathlib 기반 계층적 경로 관리 (/data/raw → /processed → /final)
* 버전 관리: GitHub - GlobalMainProject 저장소를 통해 코드 및 데이터 이력 관리.