**KOSIS 기반 추론·예측 모델**

**2025-09-24**

**0) 목표**

* **입력**: KOSIS 시계열(인구·가계·금융·가격·산업·무역 등) + 선택적 글로벌 보조지표.
* **엔진**: 상관/인과 탐색 → 구조/ML 예측 → 시나리오 충격 전파.
* **출력**: ① 거시 결론(핵심 3문장) ② 자산 전망(주식·부동산·금리·금) ③ 실행 가능한 포지셔닝/헤지 바스켓.

**1) 데이터 레이어 (ETL·표준화·버전관리)**

**스키마**

* facts\_timeseries(country, region, sector, kosis\_id, name\_ko, freq, time, value, unit, vintage\_at)
* dim\_series(kosis\_id, topic, subtopic, deflator, seasonal\_adj\_flag, transform\_hint, source\_url)

**표준화 규칙**

* 빈도 일치: 월/분기/연간 → **분기 기준 리샘플**.
* 단위/실질화: CPI/PPI로 디플레이터 적용, **per-capita / per-household / per-worker** 파생.
* 변환: 로그/차분(Δlog), 정규화(z-score), **계절조정(STL/X-13)**.
* 품질·계보: **Great Expectations**(QC), **DVC+MLflow**(데이터/모델 버전), 변환 로그 자동 각주.

**2) 관계 탐색 (Correlation Atlas)**

**파이프라인**

1. 정상성 검사(ADF/KPSS) → 필요시 차분.
2. **롤링 상관/부분상관(36~60개월)** + 다중검정(FDR).
3. **그랜저 인과성** + 크로스코릴레이션으로 **lead/lag** 추출.
4. **네트워크 그래프**(노드=지표, 엣지=유의 관계) & **Louvain 클러스터**로 테마 도출.
5. **NLG 요약**(한국어): 상위 N개 관계를 수치(ρ, p, lag) 포함 문장으로 자동 생성.

**예시 문장**  
“최근 60개월, **가계부채/GDP**는 **주택가격지수**와 **강한 양의 상관(ρ=0.62, p<0.01)**, **2~3분기 선행**.”

**3) 구조·예측 모델 (해석력+예측력 결합)**

**통계(해석력)**

* **VAR/VECM**: 공적분 검정(Johansen) 후 선택, **IRF/FEVD**로 충격 전이·설명력 분해.
* **SVAR**(부호/장기 제약): 정책금리·신용·주택가격 등 **정책/구조 쇼크** 식별.

**ML/딥러닝(예측력)**

* **XGBoost/LightGBM**: 타깃(예: GDP·주택가격·실질소비) 예측 + Feature importance.
* **LSTM/GRU 또는 TFT**: 멀티시계열 + 이벤트 피처(정책/위기 태그)까지 학습.

**앙상블**

* 구조모형 예측(설명력)과 ML 예측(정확도)을 **가중 평균/Stacking**.

**검증**

* 워크-포워드 CV, Diebold-Mariano 테스트(모형 간 예측력 비교), **팬차트(신뢰구간)**.

**4) 시나리오 시뮬레이터 (Shock → 파급)**

**입력 DSL (예시)**

name: "인구 -10%"

horizon\_quarters: 12

shocks:

population\_total: -10% step

policy\_path:

base\_rate: hold 0bp for 4Q, then -25bp for 2Q

controls:

oil\_global: baseline

usdkrw: baseline

**엔진**

* DSL→모형 입력 변환→**IRF 누적 경로** 산출(1·2차 파급: 인구→소비→고용/주택).
* 민감도 테이블: 변수별 탄력성(ΔY/ΔShock), 전이 지연.

**5) 투자 맵핑 (거시 → 자산)**

**자산군·섹터 연결**

* 주식: KOSPI/KOSDAQ & 섹터(금융·건설·필수/재량소비·수출제조 등).
* 부동산: 매매/전세지수, 거래량, 착공·분양, 리츠.
* 금리·채권: KTB 3/10년물, 신용스프레드(AA-/BBB-).
* 귀금속: 금(USD/KRW), 은(옵션).
* FX(선택): USDKRW, JPYKRW.

**탄력성 추정(초기 → 고도화)**

* 초기: 과거 **패널/단면 회귀**로 자산수익 = α + β1\*금리 + β2\*M2 + β3\*부채 + ....
* 고도화: 상태공간/시변계수(TVP-VAR), **DoubleML/IV**로 동시편의 보정.
* 결과물: **Elasticity Matrix**(거시→자산), 신뢰구간 포함.

**투자 출력**

* **전망 카드**: “금리 +50bp → 4Q 누적: 성장주 –5%, 금융 +2%, 10Y 채권 –3%, 금 +2%.”
* **포지셔닝**: 섹터 오버/언더웨이트, 롱/숏 페어, **헤지 바스켓**(예: 장기국채 숏 + 은행주 롱).
* **리스크 레이더**: 민감도 높은 이벤트(금통위, 세제 변경, LTV/DSR 조정).

**6) 리포트 & 대시보드 (자동화 산출물)**

* **Atlas**: 히트맵·네트워크·Top 관계 NLG.
* **Scenario Studio**: DSL 편집→실행→IRF/누적표/섹터 히트맵.
* **Policy & Investment**: 정책 레버(금리·LTV/DTI·세제)의 **자산 탄력성 표** + 이번 달 “3문장 요약”.
* **PDF 자동 생성**(WeasyPrint) + **웹 대시보드**(Streamlit/Dash).

**7) 품질·거버넌스**

* 데이터 테스트: 분포/결측/단위·빈도 일치, 편차 감지 알람.
* 모델 모니터링: 예측오차, 드리프트, 재학습 스케줄.
* 재현성: **MLflow**(실험/메트릭), **DVC**(데이터/피처 스냅샷).
* 컴플라이언스: 출처·버전·변환 로그 자동 각주.

**8) 백테스트(“경제적 가치” 평가)**

* **신호→포트폴리오** 백테스트(거래비용 포함), 샤프/소르티노/드로다운.
* **정책충격 과거 리플레이**(예: 금리 인상기 3회)로 모형 일관성 점검.
* **턴어라운드 탐지**: 전환점(상승→하락) 조기 경보 성능.

**9) 기술 스택 & 배포**

* 데이터/ETL: Python, pandas polars prefect, 저장 DuckDB→Postgres/BigQuery.
* 모델: statsmodels(VAR/SVAR), linearmodels(패널), xgboost/lightgbm, 선택 pytorch(LSTM/TFT).
* 서비스: FastAPI(시나리오 API), Streamlit(대시보드).
* 운영: Docker, GitHub Actions(주간 리포트 빌드), 간단 모니터링(Prometheus/Grafana 선택).

**10) 마일스톤 (MVP→프로덕션)**

**M1 (주 1–2)**

* KOSIS 100~200개 지표 ETL/표준화, QC 규칙 완료.
* Atlas v0: 롤링 상관/그랜저 + NLG 문장 1차.

**M2 (주 3–4)**

* VAR/VECM 적합 & 기본 시나리오(금리±50bp, 인구 –5%) IRF.
* 자산 탄력성(기초 회귀) 연결 → **투자 전망 카드** 생성.

**M3 (주 5–6)**

* 시나리오 DSL + 대시보드(실행/저장/비교).
* 백테스트 파이프라인(섹터 틸트/헤지 바스켓).

**M4 (주 7–8)**

* ML 앙상블 도입(TFT/XGB) + 팬차트.
* 리포트 자동 생성(PDF) + 주간 배포.

**11) 바로 필요한 입력(너가 정해주면 내가 뼈대 코드를 만들어서 붙일게)**

1. **최우선 KOSIS 지표 리스트(30~50개)**: 인구·가계부채·소득·물가·금리·주택·산업·무역 필수.
2. **자산 커버리지**: 코스피 섹터/대표 지수, 부동산(매매/전세/리츠), 채권 만기, 금(USD/KRW).
3. **정책 레버 우선순위**: 기준금리·LTV/DTI·DSR·보유/양도세·재정(주택/육아).