**분석 보고서**

1. **제주 특산물 가격 예측 AI 경진대회 분석 결과 요약**
   * **데이터 수집 및 전처리** 
     + Dataset 정보
       - Train data: 2019년 01월 01일부터 2023년 03월 03일까지의 유통된 품목의 가격 데이터
       - International\_trade data: 관련 품목 수출입 정보
       - Test data: 2023년 03월 04일부터 2023년 03월 31일까지의 데이터

○ 데이터 전처리 단계

- 시계열 특성을 학습에 반영하기 위해 timestamp 열을 월, 일, 시간 분리

- 문자열 변수를 범주형으로 라벨 인코딩

- target값인 price를 train data의 y값으로 설정

- international trade data도 위와 같은 전처리 과정을 진행

🡪 품목 데이터 값의 문자열이 있어 train데이터와 같은 라벨 인코딩 진행

🡪 무역흑자는 1로 무역적자는 0으로 한 새로운 열 생성

* + - * Train과 international trade 데이터 merge 진행
      * pd.merge(train\_x,item\_x\_cate,on=['item','year','month'],how='left')
    - 회귀분석 진행
      * 모델: sklearn.linear\_model의 LinearRegression 사용
      * X: merge data Y: price값
      * 텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

        자동 생성된 설명회귀분석 결과:

🡪 유의수준 5%내에서 무의미한 회귀계수는 없음

단, 모델의 R-squared이 0.162로 모델의 설명력이 낮음

1. **모델 설정**
   * **모델 선정**
     + - Dataset 준비

기존 merge데이터들은 무역 정보가 월별로 주어져 무역 값을 반복하기때문에 merge 데이터 셋 사용이 모델의 설명력이 증가하지 않다고 판단. 따라서 전처리를 진행한 train데이터 사용

* + - * GRU 모델
        1. 선정이유: 시계열 데이터의 예측력이 높은 모델
        2. 텍스트, 스크린샷, 그래프, 폰트이(가) 표시된 사진

           자동 생성된 설명결과: RMSE 623.49999
      * Darts 패키지 -DLinear 모델
        1. 선정이유: 시계열 데이터에 대한 예측 분석이 간편함
        2. 하이퍼파라미터 에폭과 배치사이즈 조정읕 통한 최적의 결과 예측
        3. 텍스트, 영수증, 폰트, 라인이(가) 표시된 사진

           자동 생성된 설명텍스트, 영수증, 폰트, 라인이(가) 표시된 사진

           자동 생성된 설명결과: RMSE
      * Autogluon 모델
        1. 선정이유: 시계열 데이터에 대한 예측 분석이 간편, 예측력 높음
        2. 하이퍼파라미터 조정: 내장되어 있는 함수 중 RMSE가 낮은 모델 제외
        3. 텍스트, 폰트, 스크린샷, 영수증이(가) 표시된 사진

           자동 생성된 설명결과: RMSE

1. **결론**

다양한 모델을 사용하여 가격을 예측했으며 총 915.48213의 결과를 얻음.

위의 3가지의 모델 의외에도 XGBoost, RNN, PCA, NBEATSModel, TCNMode 사용하였으나 RMSE가 컸으며 특히 XGBoost가 가장 성능이 좋지 못한 결과를 보임.

GRU모델을 사용한 결과는 좋았으나 모델의 이해 부족으로 예측 값을 추출하는 것을 성공하지 못함.

결과적으로 시계열 데이터인만큼 날짜에 관련한 전처리, 예를 들면 휴일 전처리를 하지 않아 최적의 결과를 얻지 못함. 또한 품목별로 주어진 데이터 값이 다름을 무시하고 모델링을 진행함. 예를 들어 감귤 데이터는 가격의 값이 대부분 있는 반면 상대적으로 양배추 데이터는 기록된 가격의 값이 적음.