식음업장 메뉴 수요 예측 AI 온라인 해커톤

- 링크: https://dacon.io/competitions/official/236559/overview/description
- 유형: LG Aimers | 채용 | 알고리즘 | 정형 | 시계열 | 수요 예측 | SMAPE

1. 개요
2. 데이터 구성
3. 일자별 대표 개발 및 결과
4. 결론

1. 개요
   A. 배경

> 리조트 내 식음업장은 계절, 요일, 투숙객 수, 행사 일정 등 다양한 요인에 따라 수요가 크게 변동하는 환경에 놓여 있습니다. 특히 휴양지 리조트는 단기간에 집중되는 고객 수요와 예측하기 어려운 방문 패턴으로 인해, 메뉴별 식자재 준비, 인력 배치, 재고 관리에 있어 높은 운영 난이도를 가집니다.
>
> 이러한 복잡한 운영 환경 속에서 정확한 메뉴 수요 예측은 비용 절감과 고객 만족도 향상에 있어 핵심적인 요소로 작용합니다. 최근에는 AI 기술을 활용한 수요 예측이 식음 서비스 운영의 새로운 해법으로 주목받고 있으며, 정형화된 과거 매출 데이터와 외부 요인을 함께 분석하는 방식이 빠르게 확산되고 있습니다.
>
> 이번 해커톤은 리조트 내 식음업장에서의 실전 수요 예측 문제를 AI로 해결해보는 것을 목표로 합니다. Aimers 여러분들은 실제 식음업장에서 수집된 판매 데이터를 기반으로, 각 메뉴가 1주일 동안 얼마나 판매될지를 예측하는 모델을 개발하게 됩니다. 이를 통해 데이터 기반 의사결정이 리조트 운영에 어떤 가치를 더할 수 있는지를 직접 체감할 수 있을 것입니다.

   B. 주제: 리조트 내 식음업장 메뉴별 1주일 수요 예측 AI 모델 개발 - 리조트 식음업장에서 수집된 과거의 메뉴별 판매 데이터를 기반으로, 향후 1주일간 각 메뉴의 예상 판매량을 예측하는 AI 모델을 개발

   C. 규칙
      i. 평가산식: 식음업장 별 가중치가 있는 SMAPE
         - s: 식음업장명
         - w_s: 식음업장 s의 가중치 (비공개)
         - I_s: 식음업장 s에 속한 품목 컬럼 집합
         - T_i: 품목 i에서 유효한 날짜 수 ($A_{t,i} \neq 0$)
         - $A_{t,i}, P_{t,i}$: 날짜 t, 품목 i의 실제값과 예측값

         1. 각 식음업장 별 가중치가 존재하며, '담하'와 '미라시아'는 다른 업장보다 높은 가중치로 반영됩니다. (단, 업장 별 가중치 값은 공개하지 않습니다.)
         2. 실제 매출 수량이 0인 경우에는 평가 산식 계산에 반영되지 않습니다.
         3. Public score : 전체 테스트 데이터 샘플 중 사전 샘플링된 50%
         4. Private score : 전체 테스트 데이터 샘플 100%

      ii. 외부 데이터 및 사전 학습 모델 관련 규칙
         1. 사전학습모델 사용 가능 범위
            공식적으로 가중치가 공개되었으며, 최소한 비상업적 이용이 허용된 오픈소스 라이선스(MIT, Apache 2.0 등) 하에 배포된 모델만 사용 가능합니다.

해당 조건을 만족하지 않는 모델 및 가중치는 사용할 수 없습니다.

2. API 사용 제한
   원격 서버 기반의 API 형태로만 접근 가능한 모델(OpenAI API, Gemini API 등)은 사용이 불가합니다.
   모든 모델은 로컬 환경에서 직접 실행 가능해야 하며, 외부 서버에 의존하는 방식은 제한됩니다.
3. 외부 데이터 사용 금지
   경진대회에서 제공하는 공식 데이터 외의 외부 데이터는 사용할 수 없습니다.

iii. 시계열 예측 관련 Data Leakage 방지 규칙

1. 평가 데이터는 학습에 사용할 수 없습니다.
   평가 Input(28일), Target(7일)은 어떤 경우에도 모델 학습에 활용할 수 없습니다.
   Pseudo Labeling 등 추론 결과를 이용한 재학습도 불가합니다.
   평가 데이터는 오직 추론 시점에서 입력으로만 사용할 수 있습니다.
2. 추론 시 평가 Input으로 제공된 28일 외의 데이터를 추가로 사용할 수 없습니다.
   각 평가 샘플에는 Input으로 28일간의 시계열 데이터만 제공되며, 예측 시에는 해당 구간만을 모델 입력으로 사용해야 합니다.
   평가 시점에서 Lookback 기간을 임의로 확장하거나, 추가적인 과거 데이터를 연결하여 사용하는 것은 허용되지 않습니다.
   즉, 모든 평가 샘플은 제공된 28일 데이터를 기준으로만 예측이 이루어져야 하며, 모델 구조나 데이터 처리 방식에 관계없이 28일을 초과한 입력 사용은 금지됩니다.
3. 평가 샘플은 서로 독립적으로 추론해야 합니다.
   하나의 평가 샘플 결과나 입력을 다른 샘플의 예측에 사용하는 것은 금지됩니다.
   모든 샘플은 각자의 Input(28일)만을 사용해 개별적으로 예측되어야 합니다.
4. 추론 시점 이후 정보는 사용할 수 없습니다.
   각 평가 샘플의 추론 시점은 Input의 마지막 날짜입니다.
   이 시점을 기준으로 이후의 데이터(예측 대상 포함)는 모두 미래로 간주되며 활용할 수 없습니다.
5. 외부 데이터는 사용할 수 없지만, 도메인 지식은 활용할 수 있습니다.
   대회에서 제공한 데이터 외의 외부 데이터, 크롤링, API 호출 등은 금지됩니다.
   다만, 도메인 지식 기반의 정보(예: 공휴일, 요일 등)는 활용 가능합니다. 다만, 도메인 지식 기반의 정보를 습득할 수 있는 시점도 추론 시점에 유의하여 활용할 수 있어야 합니다.
   예: "5월 5일은 공휴일이다", "일요일은 주말이다" → 추론 시점 이전에 알 수 있는 도메인 지식 기반의 정보이므로 사용 가능

2. 데이터 구성

A. train [폴더]

```
영업일자,영업장명_메뉴명,매출수량
2023-01-01,느티나무 셀프BBQ_1인 수저세트,0
2023-01-02,느티나무 셀프BBQ_1인 수저세트,0
2023-01-03,느티나무 셀프BBQ_1인 수저세트,0
2023-01-04,느티나무 셀프BBQ_1인 수저세트,0
```

i. train.csv [파일]

ii. 2023.01.01 ~ 2024.06.15의 영업장명_메뉴명별 매출 수량 정보

iii. 영업일자

iv. 영업장명_메뉴명

v. 매출수량

B. test [폴더]

```
영업일자,영업장명_메뉴명,매출수량
2025-04-27,느티나무 셀프BBQ_1인 수저세트,0
2025-04-28,느티나무 셀프BBQ_1인 수저세트,0
2025-04-29,느티나무 셀프BBQ_1인 수저세트,2
2025-04-30,느티나무 셀프BBQ_1인 수저세트,0
```

i. TEST_00.csv ~ TEST_09.csv [파일]

ii. 2025년의 특정 시점(28일)의 영업장명_메뉴명별 매출 수량 정보

iii. 영업일자

iv. 영업장명_메뉴명

v. 매출수량

C. sample_submission.csv [파일] - 제출 양식

```
영업일자,느티나무 셀프BBQ_1인 수저세트,느티나무 셀프BBQ_BBQ55(단체),"느티나무 셀프BBQ_대여료 30,000원","느티나무 셀프BBQ_대여료 60,000원"
TEST_00+1일,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
TEST_00+2일,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
TEST_00+3일,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

i. 각 영업장명_메뉴명의 TEST 파일별 +1일, +2일,..., +7일의 매출수량 예측 결과

ii. 영업일자 : TEST_00+1일, TEST_00+2일, TEST_00+3일 ... TEST_09+1일, TEST_09+2일, TEST_09+7일

3. 일자별 대표 개발 및 결과
   A. 0811_LGBM: 0.5998
      i. 팀원이 구성한 기본 LGBM 코드를 베이스로 시작
      ii. 공휴일 목록을 추가하여 피처로 입력
      iii. 테스트용 공휴일 SMAPE 함수 추가
      iv. 특성 추가로 인한 성능 저하를 확인
   B. 0812_LGBM: 0.5902
      i. Optuna 라이브러리를 적용해 파라미터 최적화 시도
      ii. 한 번 돌리는데 시간이 매우 오래 걸리기 때문에(약 3시간) 최적화 필요
   C. 0813_XGBoost: 0.5691
      i. Github 코드를 기반으로 XGBoost 기법 도입
      ii. 메뉴 별 이상치 값(상식적인 범위 밖에 있는 값)들을 IQR(사분범위) 계산을 통해 최대/최소값으로 대체
      iii. 날짜 별 주기성 각도, 래그 및 롤링 특징 생성
      iv. 예측할 날짜의 프레임 별로 래그 및 롤링 특징 계산하여 예측 수행
   D. 0815_XGBoost: 0.5324
      i. 중요도가 강조된 일부 키워드에 강제적인 가중치를 부여
      ii. '담하'와 '마라시아'에 2배의 가중치 부여
   E. 0817_XGBoost: 0.5668
      i. 가중치 부여 코드 기반으로 일부 피처 추가
      ii. 피처 추가로 인한 성능 저하를 확인
   F. 0819_XGBoost: 0.5220
      i. 2024년 데이터에 가중치 추가
      ii. 매출량이 0인 데이터에 가중치 차감
   G. 0822_XGBoost
      i. 주요 가중치 파라미터 조절
   H. 0824_XGBoost
      i. 학습 데이터 기간 내에서 무작위로 시작점을 선택하고 정해진 기간만큼을 학습하는 윈도우를 설정
      ii. 학습 데이터 윈도우 바로 다음 7일을 검증 데이터로 사용
      iii. 이 과정을 N번 반복하여 N개의 모델을 학습시키고 리스트에 저장해 최종 예측 시 N개의 모든 모델로 예측을 수행한 후 그 결과의 평균을 최종 예측값으로 사용

4. 결론
   A. 피처 엔지니어링의 효과
      i. 이상치를 IQR 값으로 대체하고 날짜 관련 파생 변수(주기성, 래그, 롤링)를 생성한 모델에서 성능이 크게 향상되는 것을 확인
      ii. 모델 자체의 성능뿐만 아니라 데이터의 특성을 잘 반영하는 피처를 생성하는 것이 예측 정확도에 매우 중요
   B. 가중치 부여의 역할
      i. 가장 뛰어난 성능을 보인 모델은 특정 키워드, 최신 데이터(2024년), 매출이 0인 데이터에 대한 가중치를 부여한 결과
      ii. 단순히 피처를 추가하는 것보다 데이터의 중요도에 따라 가중치를 조절하는 것이 성능 향상에 더 효과적
   C. 앙상블 기법
      i. 최종적으로 시도된 윈도우 기반의 앙상블 학습은 최고 기록에 근접하였음
      ii. 이를 토대로 모델을 더 정제할 시 높은 예측을 얻을 수 있을 것으로 기대됨