2024-11-04 수상작 리뷰

[제주 특산물 가격 예측 AI 경진대회]

https://dacon.io/competitions/official/236176/codeshare/9381

1. 데이터

- **데이터 전처리**: train과 test 데이터를 통합하여 timestamp를 datetime 형식으로 변환하고, 연/월/일, 요일, 주차 등을 추출하여 다양한 시계열 변수를 생성하는 과정이 포함되어 있다. 이를 통해 날짜 기반의 중요한 패턴을 추가하여 예측 성능을 높일 수 있다.
- **공휴일 변수 생성**: holidays.KR() 라이브러리를 사용해 공휴일 여부를 나타내는 변수 holiday를 추가하여, 특산물 수요와 공급에 영향을 줄 수 있는 요인을 반영하려 했다.

2. 코드 흐름

- **재현 가능한 결과**: seed_everything 함수로 난수 생성을 고정하여 실험의 일관성을 확보하고 있다.
- **모델 선택 및 결합**: XGBRegressor와 CatBoostRegressor를 사용하여 개별 모델을 학습하고, VotingRegressor로 앙상블 예측을 수행하고 있다. 다양한 모델을 결합해 예측 성능을 높이는 전략을 시도하고 있다.
- **전처리와 모델 학습 분리**: 데이터 전처리 함수와 모델 학습을 별도로 정의하여 코드의 가독성을 높였고, 개별 단계를 수정하거나 재사용하기 쉽게 구성되어 있다.

3. 배울 점

- **시간 기반 변수 생성**: 주차와 연-월 변수를 추가하는 방식, 특히 주차를 누적하여 연도 간 차이를 해소하려는 접근은 시간 데이터가 있는 예측 문제에서 유용하게 적용할 수 있는 기법이다.
- **공휴일 정보 활용**: 공휴일이 소비와 가격에 미치는 영향을 반영하는 것은 특산물 가격과 같은 도메인에 적합한 기법이다. 이는 일상적으로 고려하지 않는 요인을 모델에 포함해 성능을 높일 수 있다.
- **앙상블 모델의 활용**: 여러 모델을 결합해 일반화 성능을 향상시키려는 VotingRegressor 활용 방식은 다른 예측 문제에서도 적용 가능하며, 다양한 모델의 강점을 결합하는 기법을 배울 수 있다.