0429 cuai 스터디

**1 대회소개**

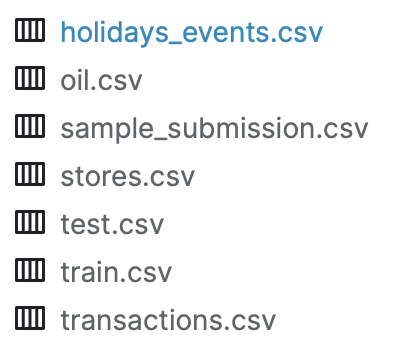
<https://www.kaggle.com/competitions/store-sales-time-series-forecasting/code>

텍스트, 화면이(가) 표시된 사진

자동 생성된 설명

Evaluation : Root Mean Squared Logarithmic Error

**2 데이터**



**train.csv :시계열로 구성된 교육 데이터입니다.**store\_nbr : 제품이 판매되는 스토어  
sales : 총 매출을 제공합니다  
프로모션: 지정된 날짜에 스토어에서 프로모션을 수행하던 제품군의 총 항목 수

**test.csv**  
테스트 데이터의 날짜는 교육 데이터의 마지막 날짜 이후 15일 동안입니다.

**stores.csv**  
도시, 주, 유형 및 클러스터를 포함한 메타데이터를 저장합니다.  
클러스터는 유사한 저장소의 그룹입니다.

**oil.csv**  
매일 기름값. 열차 및 테스트 데이터 기간 동안의 값을 모두 포함합니다. (에콰도르는 석유 의존 국가이며, 경제 상태는 유가 충격에 매우 취약합니다.)

**holidays\_csv.csv**  
휴일 및 이벤트(메타데이터 포함)

공공부문 임금은 2주에 한 번씩 15일과 그 달의 마지막 날에 지급된다. 슈퍼마켓 매출은 이것에 의해 영향을 받을 수 있다.  
2016년 4월 16일 에콰도르에서 규모 7.8의 지진이 발생했다. 사람들은 지진 후 몇 주 동안 슈퍼마켓 판매에 큰 영향을 준 물과 다른 우선 필요한 제품을 기부하는 구호 노력에 동참했다.

**3 참고할 코드**

<https://www.kaggle.com/code/andrej0marinchenko/hyperparamaters>

위 코드는 Ensemble을 이용함. (ExtraTreesRegressor , RandomForestRegressor , BaggingRegressor, VotingRegressor )

**4 공부**

**4-1 Voting 과 Bagging 차이**

Voting은 단어 뜻 그대로 투표를 통해 결정하는 방식입니다. Voting은 Bagging과 투표방식이라는 점에서 유사하지만, 다음과 같은 큰 차이점이 있습니다.

* Voting은 다른 알고리즘 model을 조합해서 사용합니다.
* Bagging은 같은 알고리즘 내에서 다른 sample 조합을 사용합니다.

즉, voting은 **서로 다른 알고리즘이 도출해 낸 결과물에 대하여 최종 투표하는 방식**을 통해 최종 결과를 선택합니다. voting은 **hard vote**와 **soft vote**로 나뉘게 되는데, hard vote는 결과물에 대한 최종 값을 투표해서 결정하고, soft vote는 최종 결과물이 나올 확률 값을 다 더해서 최종 결과물에 대한 각각의 확률을 구한 뒤 최종 값을 도출해 냅니다.

**Hard Vote**

classification을 예로 들어 보자면, 분류를 예측한 값이 1, 0, 0, 1, 1 이었다고 가정한다면 1이 3표, 0이 2표를 받았기 때문에 Hard Voting 방식에서는 1이 최종 값으로 예측을 하게 됩니다.

**Soft Vote**

soft vote 방식은 각각의 확률의 평균 값을 계산한다음에 가장 확률이 높은 값으로 확정짓게 됩니다.

가령 class 0이 나올 확률이 (0.4, 0.9, 0.9, 0.4, 0.4)이었고, class 1이 나올 확률이 (0.6, 0.1, 0.1, 0.6, 0.6) 이었다면,

class 0이 나올 최종 확률은 (0.4+0.9+0.9+0.4+0.4) / 5 = 0.44, class 1이 나올 최종 확률은 (0.6+0.1+0.1+0.6+0.6) / 5 = 0.4 가 되기 때문에 앞선 Hard Vote의 결과와는 다른 결과 값이 최종 으로 선출되게 됩니다.

**Tip**

보통 대회에서는 **Hard Vote보다는 Soft Vote 방식이 더 합리적인 방법**이라는 것이 대세입니다. 실제 성능결과도 Soft Vote 방식이 훨씬 더 잘 나와서이겠죠?

## 4-2 Blending Ensemble

Blending is an ensemble machine learning technique that uses a machine learning model to learn how to best combine the predictions from multiple contributing ensemble member models.

LNR RMSLE : 0.39

SVR RMSLE : 0.46

Mean RMSLE : 0.38

* best is Mean

LNR MAE : 53.5

SVR MAE : 106.3

Mean MAE : 70.9

* **best is LNR**

**why ?**

* **RMSLEE have higher tolerance for higher value**

**So ?**

* **Take the LNR because it has lower MAE and MAE is reliable because it’s robust to outlier : 이상값에 영향을 적게 받는 통계량**