기계학습 프로젝트 결과 보고서

김영훈   
황진익

**풀고자 하는 문제**

한국의 벼농사는 5월에 모내기를 하고 9~10월에 수확을 합니다.

모내기 이후(5월)부터 수확 이전(8월)까지의 날씨 데이터를 통해 해당 연도의 쌀 생산량을 예측하고자 합니다.

* 1차 과제에서의 기존 문제의 motivation이 부족하다는 피드백을 받고 주제를 변경하였습니다.

**사용한 데이터 셋 및 출처**

Input data: 각 지역의 년도 별 월별 평균기온, 월 합 강수량(mm), 월 합일조시간(hr) (5월~8월)

출처: 기상청(<https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36>)

Labeled data: 각 지역의 년도 별 단위면적당(10a) 쌀 생산량(kg)

출처: kosis 국가 통계 포털([시군별 논벼 생산량(정곡,92.9%) (kosis.kr)](https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1ET0034&vw_cd=MT_ZTITLE&list_id=K1_19&scrId=&seqNo=&lang_mode=ko&obj_var_id=&itm_id=&conn_path=MT_ZTITLE&path=%252FstatisticsList%252FstatisticsListIndex.do)

**사용한 알고리즘과 선택 이유**

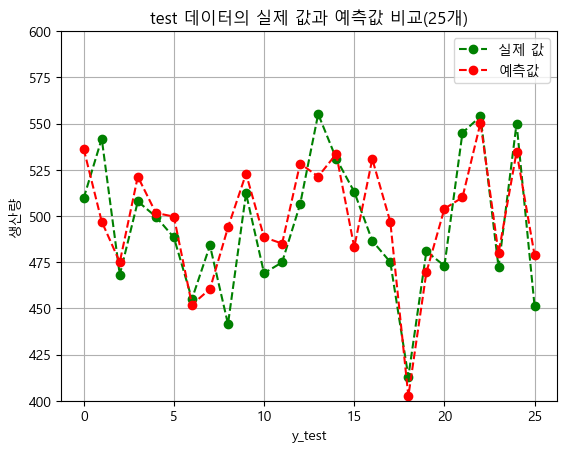
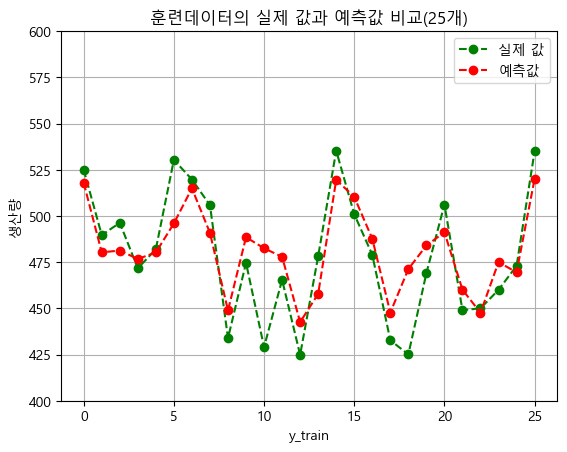
kernel method를 이용하는 SVM을 이용하였습니다. 그 이유는 다음과 같습니다

1. 날씨데이터와 쌀 생산량은 비선형적 관계를 가질 것으로 예상되므로 kernel을 적용한 SVM이 적합하다
2. 데이터 셋의 크기가 많지 않기 때문에 (약 2000개) 데이터 셋의 크기가 크지 않아도 일반화 성능이 좋은 SVM이 적합하다.

**구현 방법**

지역과 연도에 맞게 날씨 데이터셋과 쌀 생산량 데이터셋을 합쳐 훈련에 사용할 데이터셋을 만들었습니다.

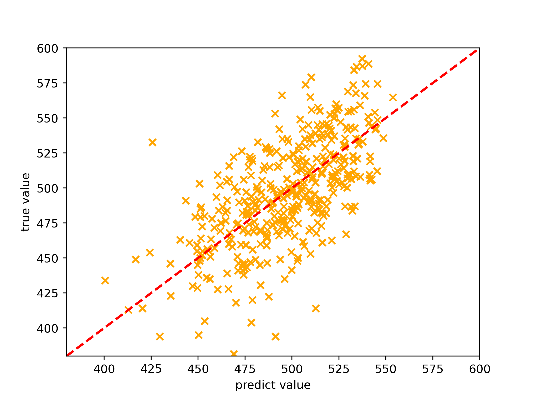
1. Input의 feature간 스케일을 맞추기 위해 StandradScale( )를 사용하여 데이터정규화 하였습니다.
2. sklearn의 SVR함수를 통해 학습을 진행하였습니다. ( training data : test data = 8 : 2 비율)  
   error는 RMSE로 계산하고 kernel은 rbf(가우시안 커널)을 적용하였습니다.  
   하이퍼파라미터는 GridSerchCv등을 이용해 튜닝하였습니다.  
   GridSerchCv는 교차검증을 지원하므로 별도의 validation 데이터를 나누지 않았습니다.  
   (최종 하이퍼파라미터 튜닝결과: 'C': 71, 'epsilon': 15, 'gamma': 0.1)  
   모델의 훈련 결과 : .   
   Test error : 29 (RMSE 기준) , 0.458 (score 기준)  
   Traing error : 20.6(RMSE 기준) , 0.753 (score 기준)
3. 훈련한 모델을 matplotlib와 shap를 이용해 분석하였습니다.



**결과 분석**

예상 결과: 5월 ~ 8월까지의 날씨데이터를 통해 쌀 생산량을 예측할 수 있다

훈련 결과: 모델이 훈련은 되었지만 생각보다 큰 error를 보였습니다.



주황색 x 마크는 test data를 활용하여 (모델이 예측한 값, 참값)을 xy 좌표상에 표시한 것입니다.

이상적인 상황에서 모델의 예측값은 모두 빨간 y=x 직선 위에 찍힙니다.

주황색 마크가 빨간색 직선과 비슷한 양상을 띠므로 모델이 쌀의 생산량을 예측하였다고 분석할 수 있습니다.

**1. 생산량을 예측한 이유**: 예상대로 날씨 데이터와 쌀 생산량의 규칙적인 연관성이 존재하며 이를 svm을 통해 학습함

**2. 예상보다 큰 error가 발생한 이유:**

1. 훈련에 사용한 데이터셋의 크기가 크지 않다
2. 사용한 데이터의 feature 이외에 생산량에 영향을 미치는 요소들이 존재한다

(ex. 병충해, 태풍 등 기상 재해의 유무, 쌀 품종, 지역별 토양 특징 등)

1. 월을 기준으로 input data를 사용하였기 때문에 해당 시기의 모든 기상 상황을 의미하지 않는다.

**3. SHAP를 통한 feature별 영향력 분석**

1) 8월에 관련한 feature가 다른 월에 관련한 feature보다 생산량에 더 큰 영향을 주었다.  
낟알이 형성되는 시기의 기후가 생산량에 많은 영향을 주는 것으로 분석할 수 있다

2) 8월은 평균기온보다 높을수록 생산량에 부정적인 영향을 미치며, 7월은 반대이다. 7월보다 8월의 평균기온이 높다는 점에서 이상적인 기온이 유지되는 것이 생산량에 긍정적인 영향을 준다고 추측된다

3) 일조시간은 모든 월을 통틀어서 높을수록 생산량에 긍정적인 영향을 주었으며,

4) 대체로 강수량이 평균보다 적을수록 생산량에 긍정적인 영향을 주었다

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

1. **기술적인 배운 점**
2. SVM에서 Hyper-parameter따라 성능이 변화하고 또 overfitting이 발생하는 것을 확인하는 등   
   Hyper-parameter가 모델의 결과에 주는 영향을 튜닝하는 과정에서 확인해볼 수 있었습니다.
3. 데이터에 따라 데이터정규화를 하는 것이 모델의 성능을 향상시킬 수 있는 것을 확인할 수 있었습니다.