안녕하세요 ~조 박수민입니다

제가 구현한 모델은 최대한 간단하게 구현하고, 학습시간을 줄이면서, 비선형적으로 만드는 방법을 생각했는데요

Ppt에 언급하지는 않았지만, 저희 조원 대부분 공통으로 사용한 전처리(scaler나, 결측치 평균치 보간) + PCA+Polynomial Feature+Ridge로 구현하였습니다.

다차항을 쓰되, 과적합은 Ridge를 해결하는 것이 기본 아이디어 골자입니다.

최대한 비선형적으로 만들기 위해 Polynomial Feature 기법을 사용했지만, 다변수 다차항까지 고려한다면 가령 3차만 하더라도 x^2y, xyz, y^2z 처럼 feature의 수가 너무 대폭 늘어나는 것을 확인할 수 있습니다. 이는 과적합, 즉 성능 감소로 이어지죠. 그렇기에 미세조정을 하거나 특성 선택의 효과가 있는 모델, 가령 lasso등을 사용하는 것이 좋을 것이라 생각하고 이것저것 방안들을 고려해 보았습니다.

생각해본 방안들 중, 최종적으로는 성능이 가장 좋았던 것은 Pca로 우선적으로 feature수를 감소시키고, Polynomial Feature를 하여 비교적 최종 feature수를 크게 감소시킨 다음 Ridge로 과적합을 하는 방향이 가장 좋았습니다.

전처리, PCA, Polynomial Feature, Ridge로 구현하였으며 이때 Validation 데이터셋에서 RSME 모델 평가를 해본 결과 PCA의 사용 특성수는 7개, Polynomial Feature 2차항 사용, Rideg의 alpha=5를 했을 때의 성능이 가장 좋았습니다.

마지막 모델 평가는 큰 데이터 셋에서 Validation RMSE는 46이 나왔고, 작은 데이터셋에서 학습 후 test데이터 셋으로 평가를 해 본 결과도 46으로, 다른 분들에 비하면 성능이 조금 아쉬운 수준에 그쳤습니다.