안녕하세요 12조 소프트웨어학과 박수민입니다.

저희 조는 다른 분들이 조를 만들고 남은, 같이 조를 만들 사람이 없는 사람들이 모여서 만든 조인데요, 서로 안면식이 없는 만큼 호흡을 맞추는데 비교적 많은 노력이 들것이라 생각했습니다. 이에 역발상으로 조원 총 5명이 각자 개인플레이로 자신의 모델을 구현해보고, 평가하고, 이를 앙상블로 결합하여 결과를 확인해보는 식으로 진행하였습니다.

발표 개요입니다. 선정 데이터 소개와, 앞서 설명했던, 저희 조원들의 모델소개 및 평가, 그리고 앙상블 모델의 소개 및 평가로 이루어져있습니다.

선정 데이터는, 헤커톤에서 선정한 주제인 “서울시 따릉이 대여량 예측”를 그대로 사용하였고, 다만 데이터셋은 접근권한 문제로 유사한 데이터셋을 따로 구해서 사용하였습니다.

파라미터들은 다음과 같습니다. 저희 모델에서 Id는 학습에서 제외하였고, 대여시간대, 1시간전 온도, 강수량, 풍속, 습도, 가시거리, 오존농도, 미세먼지 농도, 초미세먼지농도를 통해 대여량을 예측하는 모델입니다.

첫번째는 제가 구현한 모델입니다. 저는 학습 속도, 직관성 등을 고려해서 최대한 간단하게 구현하고, 학습시간을 줄이면서, 비선형적으로 만드는 방법을 생각했는데요

이부분은 Ppt에서 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 있었습니다. 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 과적합 방지를 하는 방향이 가장 좋았습니다.

최종적으로 전처리, PCA, Polynomial Feature, Ridge로 구현하였으며 이때 Validation 데이터셋에서 RMSE 모델 평가를 해본 결과 PCA의 사용 특성수는 7개, Polynomial Feature 2차항 사용, Rideg의 alpha=5를 했을 때의 성능이 가장 좋았습니다. 막 10차항 사용에 alpha = 50000 성능이 더 좋게 나올 때도 있지만 튈 때도 있더라고요. 학습시간도 오래걸려서 의도와 맞지않던 부분도 있었고요

마지막 모델 평가입니다. Validation RMSE는 46이 나왔고, 다른 분들에 비하면 성능이 조금 아쉬운 수준에 그쳤습니다.