최종 결과 보고서 – 4조

데이터셋으로 2012-01-01 ~ 2015-12-31 까지의 미국 시애틀의 날씨를 사용했다.

데이터셋 안에는 각 날짜의 강수량, 최고 기온, 최저 기온, 풍속, 그리고 날씨(예를 들어 rain, snow 등등)가 들어있다.

다양한 입력변수들을 주고 출력 변수로는 그 날이 화창한 날인지(sunny = 1) 그렇지 않은지(not sunny = 0) 분류하는 모델을 만들어봤다.

분류 모델로는 랜덤 포레스트 모델을 사용하였고, 기본적으로 주어진 데이터들과 화창한 날에 대한 자가상관함수를 구해서 입력변수들로 넣어준 다음 얻은 모델 평가 지표들은 다음과 같다.

Random Forest - Accuracy: 0.63

Random Forest - Precision: 0.55

Random Forest - Recall: 0.54

Random Forest - F1 Score: 0.54

다음로는 강수량, 최고 기온, 최저 기온, 풍속에 대해 3일 이동 평균을 구해서 입력 변수로 넣어주었다. 이 변수들로 얻은 모델 평가 지표들은 다음과 같다.

Random Forest - Accuracy: 0.74

Random Forest - Precision: 0.78

Random Forest - Recall: 0.54

Random Forest - F1 Score: 0.63

두 번째 모델에서 정확도와 정밀도가 상승한 모습을 보여주었지만, 재현율이 여전히 그대로이다. 따라서 재현율을 높이기 위해 XGBoost기법을 사용해보았다. XGBoost기법을 통해 얻은 모델 평가 지표들은 다음과 같다.

XGBoost- Accuracy: 0.71

XGBoost- Precision: 0.69

XGBoost- Recall: 0.65

XGBoost- F1 Score: 0.65

재현율이 0.11 상승하였지만 정확도와 정밀도가 떨어진 모습이다. 마지막으로 그리드 서치를 통해 최적의 모델을 찾아보고자 한다. scoring='f1'는 f1점수를 기준으로 최적화해 가장 좋은 모델을 서치해준다. 그리드 서치를 통해 얻은 모델 평가 지표들은 다음과 같다.

Optimized Random Forest - Accuracy: 0.74

Optimized Random Forest - Precision: 0.81

Optimized Random Forest - Recall: 0.54

Optimized Random Forest - F1 Score: 0.63

프로젝트를 진행하면서 배운점 및 느낀점

-김재훈 : 재현율을 높이려고 하니 오히려 정확도와 정밀도가 낮아지는 모습을 보고 적절한 하이퍼파라미터 조절과 충분한 데이터셋이 중요하다는 사실을 알게됐다. 하면서 솔직히 무슨 내용인지는 정확하게 이해가 되진 않았지만, github에 관해서도 배우고 기본적인 인공지능 모델에 대해 배운 것이 나중에 전공을 배울 때 도움이 될 것 같다.

-남태오 : 데이터를 제대로 탐구하고 이해하는 작업 없이는 성공적인 모델링과 유의미한 예측이 어렵다는 것을 느꼈기에 데이터 이해와 준비과정이 정말 중요한 부분이란걸 배웠다.

-박시은 : kaggle로 머신러닝 코드를 둘러보고 파이썬으로 바꿔보는 과정이 유익했다.

-성승훈 : 여러 가지 파이썬 모듈을 통해 머신러닝을 할 수 있다는 것과 f1 점수를 통해 다양한 머신러닝 모듈 중에 어떤 걸 사용하느냐에 따라 달라지는게 신기하였고 kaggle을 통해 다양한 데이터를 얻을 수 있다는게 좋았다.

-이수연 : 코드를 수정하면서 실행해보니 다양한 결과값이 나오는 게 신기하고 스스로 코드를 작성해서 실행해보고 싶다는 생각이 들었다. 머신러닝 모델을 찾아보면서 머신러닝에 대한 공부를 더 열심히 해야겠다고 다짐했다.