안녕하세요 녹조 예측 모델링을 주제로 발표할 AI 18기 이민엽입니다.

오늘 제가 말씀드릴 주제는 녹조 현상에 대한 예측입니다. 녹조 현상의 원인으로는 첫번째 오염물질이 유입입니다. 오염물질이 유입되면 수중 질소와 인의 농도가 높아져 조류가 생장하기 좋은 환경으로 변합니다. 두번째 원인으로는 수온의 상승입니다. 모두 아시겠지만, 추운 환경보다는 더운 환경에서 대부분의 생명체가 활성화하게 됩니다. 따라서 녹조 역시 수온이 증가하면 더 높은 활성도를 보입니다. 다른 조건으로는 일사량이 증가하고, 물순환이 정체될 때 녹조 현상이 일어나게 됩니다. 그렇다면 녹조 현상은 어떤 문제를 일으킬까요? 녹조가 표면에 가득 차게되면 광합성을 하기 위해 햇빛을 사용하게 되고, 자연스럽게 수중에 일사량이 감소하게 됩니다. 물 속에 있는 광합성 식물들은 광합성을 하지 못해 산소를 소비하게됩니다. 이에 따라 수중 산소량이 감소하게 되고, 산소로 호흡하는 어류나 기타 동물들은 산소 결핍을 겪게 됩니다. 따라서 녹조 현상이 심화되면 동식물의 시체가 부패하고, 부패하는 과정에서 많은 미생물과 세균이 산소를 이용하여 분해하기 때문에 악순환의 반복이 일어납니다. 결과적으로 녹조 현상은 수생태계에 지대한 악영향을 미치게 됩니다. 많은 연구에서 주로 이야기하는 유해 남조류는 마이크로시스틴, 수다나베나, 오실라토이얼, 아파니조메논을 유해 남조류로 분류합니다. 저는 이번 프로젝트에서 이 네 종류의 유해 남조류를 이용했습니다.

제가 이 프로젝트를 하면서 생각했던 가설에 대해 이야기하겠습니다. 첫번째 가설로는 일반적으로 많은 미생물들은 커뮤니티를 이루고 있습니다. 따라서 우리가 타겟으로 보는 네 종류의 유해 남조류 역시 커뮤니티를 이룰 것이고, 이를 통해 예측하는 모델을 만들 수 있다고 생각했습니다. 두번째 가설로는 녹조의 생존 환경은 물이므로, 수질 지표에 영향을 받는다고 생각했습니다. 따라서 수질 지표를 이용해서 유해 남조류를 예측할 수 있다고 생각했습니다. 이 두 가설을 가지고 이번 프로젝트에 참여했습니다.

다음으로는 제가 사용한 데이터와 그 데이터의 전치리에 대해 말씀드리겠습니다.

제가 사용한 정보는 연세대학교 환경에너지공학부의 생태환경생명공학연구실의 자료를 이용했습니다. 이번 데이터의 컬럼들은 다음과 같습니다. 추가적으로 설명 드릴 부분은 제가 말씀드린 유해 남조류인데요. 제가 초반에 녹조의 위험성에 대해 넓게 설명드렸지만, 이번에는 저희가 볼 타겟에 대한 추가 정보를 이야기해드리겠습니다. 크로시스틴은 급격히 증식할 때 많은 독소물을 배출합니다. 수다나베나 역시 대량 번식 시 독소를 생성하여 수질 오염을 일으킵니다 오실라토이얼은 번식을 하여 농도가 높아지면 수질 오염 관련 문제를 일으킵니다 아파니조메논은 농도가 높아질수록 일사량을 감소시켜 햇빛의 침투를 막습니다.

분석하기 전에 진행한 데이터 전처리에 대해 말씀드리겠습니다. 우선 Rain column의 이상치를 제거했습니다. 비가 많이 물이 빠르게 순환하고, 수온이 내려가게 됩니다. 따라서 녹조를 분석하는데 필요가 없기 때문에 제거했습니다. 두번째로는 위치 정보 처리입니다. Lake Column은 독립된 호수가 아니라 서로 연결된 호수입니다. 대략적인 지도를 보시면 상류와 하류가 나눠진 것을 확인할 수 있습니다. 따라서 서로 영향을 미치기 때문에 추가 전처리를 했습니다. 전처리 과정은 보이시는 것과 같이 숫자로 전환시켰습니다.

세번째로 결측치를 제거했습니다. 결측이 있는 행이 2개가 있었습니다. 해당 행은 모든 열에 대해서 결측이었기 때문에 dropna를 이용하여 행 자체를 제거했습니다. 마지막으로 계절에 대해서 전처리를 진행했습니다. 보시는 것처럼 봄에 유해 남조류 4종의 양이 적습니다. 이는 모델링을 할때 오차가 발생하기 때문에 제거했습니다.

첫번째 가설입니다. 유해 남조류를 커뮤니티로 예측하고자 모델링을 진행했습니다. 이번 가설에서 중점적으로 보셔야할 부분은 시계열 데이터에 대한 클러스터링입니다.

시계열 데이터를 어떻게 클러스터링을 했을까 라고 궁금하신 분들이 많으실텐데요 저는 동적시간와핑, 즉 다이나믹 타임 와핑을 이용했습니다. 그림에 보시는 것처럼 시간과 상관없이 두 연속적인 데이터 간의 유사성을 측정하는 알고리즘입니다. 이 알고리즘의 장점은 두 시계열을 매칭시킬 경우에 사용하며, 유클리디안 거리 방법과 달리 왜곡과 변형된 파형에 대해서도 적절히 매칭해줍니다. 즉 연속된 두 시계열 데이터가 서로 다른 시간에 있어도, 유사하다면 두 데이터 사이의 거리 짧다고 측정됩니다. 따라서 DTW의 값이 작을수록 두 데이터는 유사하다고 해석할 수 있습니다.

저는 이 알고리즘 사용하여 유해 남조류와 유사한 거동을 가지는 미생물들을 분류했습니다. 첫번째 그래프르 보시면 빨간색으로 그려진 것이 마이크로시스틴의 거동이고, 뒤에 회색으로 보이는 것이 유사성 기준 상위 5개를 그렸습니다. 보시면서 어? 비슷하지 않은가 같은데? 라고 생각하실 수도 있는데요. 이 데이터의 행은 475개로, 현재는 150개부터 450개까지만 그렸습니다. 이 범위 외에서도 유사한 거동을 보일 수 있는 점을 참고해주시면 좋을 거 같아요 오른쪽 그래프는 마이크로 시스틴과 각 미생물 사이의 거리입니다. 두 그래프 사이의 거리가 가까울 수록 유사하다고 해석할 수 있습니다. 따라서 보이시는 5개의 미생물이 가장 유사하다고 볼 수 있습니다.

해당 미생물과 수질 데이터를 이용하여 분석을 진행했습니다. K가 45라는 의미는 상위 45개의 미생물을 이용했다는 의미이고, 왼쪽 그래프는 K의 값에 따라 변하는 R2 score를 시각화한 그래프입니다. 파란색은 트레인 데이터고, 주황색은 벨리데이션 데이터입니다. R2 socre은 벨리데이션 기준 0.1로 매우 낮게 나왔습니다. 또한 순열 중요도를 확인했을 때 비슷한 거동을 보인다고 분석한 미생물들 중 두개만 상위 10개 안에 포함되었습니다.

다음으로는 수다나베나에 대해 동일하게 진행했습니다. 두 그래프 역시 마이크로 시스틴과 동일한 의미를 가지고 있습니다

수나다베나의 경우 벨리데이션 기준 R2 score이 0.23이 측정되었습니다. 또한 비슷한 거동을 지닌다고 분석된 상위 5개의 미생물들 중 어떤 것도 순열 중요도에 포함되지 않았습니다.

오실라토이얼 역시 동일한 의미의 그래프입니다.

오실라토이얼의 R2 score은 벨리데이션 기준 0.12입니다. 이 역시 매우 낮은 것을 확인할 수 있습니다. 심지어 하이퍼파라미터를 최적화 했음에도 과적합이 발생한 것을 확인할 수 있습니다. 오른쪽 그래프를 보시면 비슷한 거동을 가진 상위 5개 미생물들 모두 순열 중요도상위 10개에 포함되지 않았습니다.

아파니조메논 역시 동일한 의미의 그래프입니다.

아파니조메논의 R2 score은 0.01로 매우 낮았습니다. 또한 순열중요도 역시 상위 2개만 의미가 있었으며, 비슷한 거동을 가진 5개의 미생물들은 상위 10개의 순열중요도에 포함되지 않았습니다.

Test set에 대한 R2 score은 의미가 없다고 판단되어 진행하지 않았습니다. 왜 R2 score이 낮게 나왔을까요? 우선 미생물의 커뮤니티는 단순히 증감으로 확인할 수 없습니다. 미생물의 커뮤니티는 굉장히 복잡하고 상호 관계적이며, 단순히 연관이 되어있다 로 끝나지 않고 포식 및 기생 공생 등 다양한 관계가 존재합니다. 그렇기 때문에 증감 분석 만으로 미생물 커뮤티니를 분석하기에는 한계가 있었습니다. 두번째로 환경 자료 특성상 정확도가 떨어집니다. 미생물을 측정하기 위해서는 현미경을 이용하여 직접 카운팅을 하거나, 프로그램을 이용하여 카운팅을 진행합니다. 하지만, 넓은 범위의 환경 분석은 어렵고, 모집단에 대한 추론이 어렵습니다. 그렇기 때문에 이번 모델의 R2 Score이 낮게 나왔다고 판단했습니다. 이번 가설에서는 다이나믹 타임 워핑을 이용한 미생물 거동 분석을 진행했고, 비슷한 거동의 미생물끼리 클러스터링을 진행했습니다. 물론 비지도학습의 클러스터링과는 다르지만, 대략적인 집단을 파악 수 있었습니다. 결과적으로 제가 세운 가설이 잘못됐을 수도, 혹은 더 전문적인 커뮤니티 분석이 필요하다고 생각합니다.

두번째 가설로는 미생물이 거주하는 수질에 대한 지표를 통해 녹조 현상 정도 예측을 진행해보았습니다.

제가 처음에 말씀드린 네 종류의 유해 남조류로 클러스터링을 진행했습니다. 4차원 데이터에 대해 PCA를 사용했고, 2차원으로 시각화했습니다. 오른쪽 그래프는 주성분에 따른 분산을 나타내었고, PC를 2개 사용하는 것은 정보를 잘 담고 있다고 말할 수 있습니다. 저는 클러스터를 4개로 분류했습니다. 분류 기준은 엘보우 메소드를 사용했습니다. 결과는 왼쪽 그래프와 같습니다. 왼쪽 그래프를 보시면 노란색 점을 확인할 수 있으신대요, 저 점은 데이터가 1개이므로 분류 모델링에 오차를 만들 수 있다고 판단하여 제거했습니다. 따라서 3종류의 클러스터를 가지고 분류 분석을 진행했습니다.

다음 보실 그래프는 각 유해 남조류 별 클러스터의 분포입니다. 마이크로시스틴은 분산이 정확히 나눠져 있어 크기를 비교할 수 있었습니다. 크기는 클러스터 1이 가장 컸고, 그 다음으로 2, 0 순서였습니다. 수다나베나는 보시는 것처럼 분산의 크기를 비교할 수 없어 크기 분석이 불가능했습니다.

오실라토이얼은 클러스터 1 평균이 가장 컸고, 나머지 두 클러스터는 비교가 불가능했습니다. 아파니조메논은 세 클러스터를 비교할 수 없었습니다. 결과적으로 클러스터 1은 녹조 현상이 심각한 상태, 클러스터 2는 녹조 현상이 의심스러운 상태, 클러스터 0은 평상시라고 할 수 있습니다.

마지막으로 해당 모델에 대한 컴퓨전 메트릭스입니다. 컨퓨젼 메트릭스를 보시면 대부분의 값들이 0에 쏠려 있는 것을 확인할 수 있습니다.

이 모델링의 결론입니다. 해당 모델의 예측 정확도가 0.97로 나왔습니다. 하지만, 가장 많은 빈도로 만든 baseline의 예측 정확도 역시 0.94였습니다. 결과가 이렇게 나온 이유는 클러스터 0의 데이터가 압도적으로 많았기 때문입니다. 이를 위해서는 추가적인 자료 수입이 필요하며, 이 모델은 적합하지 않습니다. 추가적인 해야할 부분은 클러스터링입니다. 트레인과 테스트 셋의 클러스터는 동일하게 적용하면 안 됩니다. 물론 해당 데이터를 학습시킬때 정보누수를 막기 위해 클러스터에 사용된 컬럼을 제거했습니다. 하지만 기존 방법은 트레인 데이터 셋은 파이썬으로 클러스터링을 한 후, 테스트 데이터 셋에는 전문가가 직접 클러스터링을 진행합니다. 이렇게 생긴 클러스터와 예측된 클러스터를 비교하여 모델의 성능을 평가합니다. 따라서 해당 방법이 잘못되었지만, 녹조 분류 인사이트를 얻을 수 있었습니다.

이번 프로젝트의 최종 결론입니다. 가설 1은 커뮤티니 분석을 이용해 녹조 발생량을 예측하고자 했습니다. 하지만 커뮤니티 분석이 단순 증감으로는 파악하기 어렵고 더 자세한 조사가 필요하다는 한계가 있었습니다. 가설 2는 수질 데이터를 이용해서 녹조 정도를 예측하려고 했지만, 대부분의 클러스터가 0으로 쏠러 있어 정확한 모델링을 하기에는 어려웠습니다. 또한 테스트와 트레인 데이터셋에 따로 클러스터링 진행을 해야 합니다. 이번 프로젝트에서 높은 성능을 보이는 모델을 만들지는 못했지만, section 1의 기법들과 혼합해 사용해보았고, 앞으로는 더 많은 방법을 이용하여 인사이트를 넓혀나가고싶습니다. 감사합니다.

지금까지 녹조 발생량 예측 모델을 발표한 이민엽입니다. 감사합니다.